Mathematical and numerical analysis of models of condensed-matter physics

Antoine Levitt

Université Paris-Est

Habilitation à diriger des recherches

Présentée par
Antoine Levitt
MATHERIALS, Inria Paris, CERMICS, École des Ponts

Mathematical and numerical analysis of models of condensed-matter physics

Soutenue le 2 octobre 2020

Rapporteurs :	**Stefan Goedecker** Université de Bâle **Patrick Joly** Inria Saclay, ENSTA **Stefan Teufel** Université de Tübigen
Examinateurs :	**Grégoire Allaire** École Polytechnique **Éric Cancès** École des Ponts **Laura Grigori** Inria Paris **Nicola Marzari** EPFL **Éric Séré** Université Paris Dauphine

Contents

Remerciements

Je remercie tout d'abord les membres du jury, et en particulier les rapporteurs, pour leur travail. Mention spéciale à Patrick Joly : ses efforts sur le manuscrit m'ont amené à revisiter et (j'espère) clarifier certaines parties.

Ce manuscrit a bénéficié de la relecture et des commentaires de mes collègues de MATHERIALS Claude, Eric, Gabriel, Tony and Virginie, que je remercie chaleureusement.

Au cours de ces quelques années à Dauphine, au CEA, à Jussieu et au CERMICS, j'ai eu la chance de côtoyer et de travailler avec des personnes de grande qualité humaine et scientifique. Les lister ici serait trop long, mais je les remercie tous. Je remercie en particulier les diverses personnes qui m'ont pris sous leur aile et m'ont appris mon métier, mes co-auteurs pour leur travail et leurs perspectives différentes de la mienne, et les doctorants et post-doctorants avec qui j'ai eu la chance de travailler.

Je remercie en particulier les personnes qui m'ont mis le pied à l'étrier à mon arrivée à MATHERIALS, et notamment Éric Cancès pour m'avoir confié des responsabilités d'encadrement, Claude Le Bris pour avoir été un chef d'équipe exemplaire et Gabriel Stoltz pour l'enseignement.

Merci à tous les travailleurs de l'ombre, sans qui mon travail n'aurait pas été possible : les contribuables qui paient mon salaire, les personnels administratifs, et les développeurs des logiciels libres que j'utilise (et notamment Emacs, Julia et leurs écosystèmes).

Merci à Michael Herbst pour s'être lancé avec moi dans l'aventure DFTK, et à tous ceux qui y ont contribué directement ou indirectement.

Ce manuscrit est dédié à ma famille, à Adrienne et Lucie, et aux absents.

CHAPTER I

Results obtained and publications

My research has focused on the mathematical and numerical analysis of methods for electronic structure, with a particular focus towards the periodic systems that form the basis of our understanding of solids. In Chapter II, I give a brief introduction to the theory of electronic structure, and in Chapters III, IV and V I outline my contributions respectively to the analysis of periodic systems, to the construction of Wannier functions and to the design of iterative numerical schemes.

Here I briefly summarize my contributions, and the resulting publications. All these contributions have been performed after my PhD thesis. In this chapter, my papers are cited as a letter and a number, where the letter refers to the chapter (P, W and I for chapters III, IV and V respectively). In the rest of the document, the bibliography is split between my papers, cited as a number [**1**], and other papers, cited as [Abc1].

I.1. Periodic quantum systems

In Chapter III, I study the theory and computation of properties of periodic quantum systems. The goal is to provide rigorous mathematical models for the ground state and response properties of solids, and to analyze the numerical methods used to compute them. I have in particular focused on the case of metals, which present distinct theoretical and numerical challenges due to the absence of a gap between occupied and virtual states.

In [P1], we have studied the convergence properties of commonly used numerical methods for metallic systems: direct Brillouin zone integration and smearing methods. We proved error estimates which provide practical guidance on the choice of numerical parameters. In [P2], I have studied the existence theory and screening properties of finite-temperature solids, and proved that an appropriately preconditioned self-consistent field method has a convergence rate independent of the size of the supercell. In [P3], we have studied a simple quantum model of electronic transport in crystals, and proved in a unified setting estimates for the current in insulators, metals and semimetals, in the linear response and adiabatic regime.

My co-authors on this topic are Eric Cancès, Virginie Ehrlacher, Sami Siraj-Dine (ENPC and Inria), David Gontier (Dauphine), Clotilde Fermanian Kammerer (Créteil), and Damiano Lombardi (Inria Paris). The methodological tools used are elliptic and evolution partial differential equations, nonlinear analysis, perturbation theory, numerical analysis and complex analysis.

[P1] E. Cancès, V. Ehrlacher, D. Gontier, A. Levitt, and D. Lombardi. Numerical quadrature in the Brillouin zone for periodic Schrödinger operators. Accepted in *Numerische Mathematik*, 2019.

[P2] A. Levitt. Screening in the finite-temperature reduced Hartree-Fock model. Accepted in *Archive for Rational Mechanics and Applications*, 2018.

[P3] E. Cancès, C. Fermanian Kammerer, A. Levitt, S. Siraj-Dine. Coherent electronic transport in periodic crystals. *Submitted*, 2020.

I.2. Wannier functions

In Chapter IV, I study the construction and properties of Wannier functions for periodic systems. Wannier functions can be dually seen in real space as a localized representation of an energy subspace, or in reciprocal space as a method for the interpolation of band structures. They prove useful in a large number of contexts, from understanding bonding in crystals to computing electron-phonon interactions. Mathematically, the very existence of localized Wannier functions is highly non-trivial, failing in the case of materials with topological order. Numerically, this means that the standard methods are fragile, depending sensitively on user input. This problem is more pronounced in the case of metals, whose band structure is entangled.

In [W1], we have proposed and implemented a method that robustly builds localized Wannier functions for isolated band structures, and extended it to topologically challenging systems in [W4]. In [W2], we proposed a mathematical definition of Wannier functions for entangled band structures, mirroring established numerical practice, and proved their localization under certain conditions. We finally proposed new numerical methods for their construction in [W3], and showed that the most widely used class of Wannier functions for entangled band structures are only algebraically localized.

My co-authors on this topic are Eric Cancès, Sami Siraj-Dine, Gabriel Stoltz (ENPC and Inria), Horia Cornean (Aalborg), Anil Damle (Cornell), David Gontier (Dauphine), Lin Lin (Berkeley), Domenico Monaco (Rome), Gianluca Panati (Rome). The methodological tools used are differential geometry and homotopy theory.

[W1] E. Cancès, A. Levitt, G. Panati, and G. Stoltz. Robust determination of maximally localized Wannier functions. *Physical Review B*, 95(7):075114, 2017.

[W2] H.D. Cornean, D. Gontier, A. Levitt, and D. Monaco. Localised Wannier functions in metallic systems. In *Annales Henri Poincaré*, pages 1–25. Springer, 2017.

[W3] A. Damle, A. Levitt, and L. Lin. Variational formulation for Wannier functions with entangled band structure. *Multiscale Modeling & Simulation*, 17(1):167–191, 2019.

[W4] D. Gontier, A. Levitt, and S. Siraj-Dine. Numerical construction of Wannier functions through homotopy. *Journal of Mathematical Physics*, 60(3):031901, 2019.

I.3. Iterative methods

Chapter V gathers collaborative works on iterative numerical methods in different contexts (electronic structure, polarizable force fields, exploration of potential energy surfaces, Bose-Einstein condensation).

In [I1,I2], based on my postdoctoral work at CEA, we proposed a Chebyshev method to avoid communications in the iterative diagonalization of the Hamiltonian of plane-wave density functional theory, yielding large speedups on massively parallel supercomputers. In [I5], we studied the convergence and non-convergence properties of the widely used dimer method to compute saddle points of potential energy surfaces, concluding that such methods intrinsically lack robustness. In [I4], we proposed a constrained preconditioned conjugate gradient for the computation of Bose-Einstein condensates, more efficient than the previously used methods based on imaginary time integration (gradient descent). In [I3], we differentiated explicitly a truncated conjugate gradient to compute efficiently exact forces for polarizable force fields, leading to more stable dynamics.

My co-authors on this topic are Xavier Antoine, Qinglin Tang (Nancy), Christoph Ortner (Warwick), Qinglin Tang, Marc Torrent (CEA), as well as a large body of authors around the Tinker-HP and ABINIT codes. The methodological tools used are linear and nonlinear iterative methods, numerical optimization and high-performance computing.

[I1] A. Levitt and M. Torrent. Parallel eigensolvers in plane-wave density functional theory. *Computer Physics Communications*, 187:98–105, 2015.

[I2] X. Gonze, F. Jollet, et al. Recent developments in the ABINIT software package. *Computer Physics Communications*, 205:106–131, 2016.

[I3] F. Aviat, A. Levitt, B. Stamm, Y. Maday, P. Ren, J.W. Ponder, L. Lagardere, and J-P. Piquemal. Truncated conjugate gradient: an optimal strategy for the analytical evaluation of the many-body polarization energy and forces in molecular simulations. *Journal of Chemical Theory and Computation*, 13(1):180–190, 2016.

[I4] X. Antoine, A. Levitt, and Q. Tang. Efficient spectral computation of the stationary states of rotating Bose–Einstein condensates by preconditioned nonlinear conjugate gradient methods. *Journal of Computational Physics*, 343:92–109, 2017.

[I5] A. Levitt and C. Ortner. Convergence and cycling in walker-type saddle search algorithms. *SIAM Journal on Numerical Analysis*, 55(5):2204–2227, 2017.

CHAPTER II

Introduction to electronic structure

The ultimate goal of molecular simulation is to compute from first principles the properties of molecular systems, from a single atom in vacuum to macroscopic materials. This is certainly an ambitious program. First, while the basic equations dictating the behavior of microscopic particles have been known since the 1920s, their complexity grows exponentially with the system size, making it impossible to use them directly. Second, this is a multiscale problem, with length and time scales spanning more than ten orders of magnitude. Fortunately, both these difficulties are amenable to approximations that make them tractable. Theoretical and methodological improvements, combined with the continual increase in computing power, have resulted in molecular simulation transitioning from a theoretical tool yielding insight into the fundamental behavior of matter to a quantitative theory, able to predict many properties to good accuracy. As a result, answering the following questions from first principles is now routine, and can be computed in a matter of minutes on a laptop:

- What is the equilibrium geometry of the H_2O molecule?
- What is the reaction rate of the reaction $H_2 + F_2 \to 2\ HF$?
- What is the lattice structure of silicon?
- What is the color of gold?

More complicated properties such as phase diagrams or electrical conductivities can be obtained with more computer power.

Accessing more complicated properties and systems is an extremely active area of research, with direct applications in chemistry, biology, physics and materials science. Contemporary methodological research focuses on large systems, correlated electrons, and interactions of systems with their environment.

The field of molecular simulation forms a rich playground for applied mathematics. The first challenge is theoretical. Are the equations well-posed? What are the properties of their solutions? How do they behave in various regimes? Understanding the Schrödinger equation and its consequences has historically led to the development of large parts of mathematics, such as functional analysis and spectral theory. Since then, molecular simulation in general and electronic structure in particular have proven to be rich sources of mathematical problems, as well as avid consumers of mathematical tools: very few mathematical fields do not apply in one form or the other to the problems of molecular simulation.

The second challenge is numerical. Once an appropriate approximation and a system of interest have been selected, how do we solve the equations robustly, accurately and efficiently? The numerical computation of electronic structure has a long history in both quantum chemistry and solid-state physics. Starting from theories already established in the '60s, methodological advances and the growth in computing power have pushed electronic structure from a qualitative theory of fundamental physics to a commonly-used tool able to make quantative predictions on systems of scientific and industrial importance. The rise of new methodologies such as high-throughput screening of materials, as well as changes in computing paradigms towards mass parallelism, impose ever more constraints on the numerical methods used.

The purpose of this introduction is to give an overview of the methodology for the simulation of molecules and materials at the quantum level. This methodology is of interest both to chemistry and biology, which simulates molecules and proteins in their environment, and to condensed-matter physics, which simulates crystals and their defects as well as liquids. The equations to be solved are the same; however, the environments, properties of interest, approximations, discretizations and solution methods may not be. In practice, both communities are separate, with relatively little overlap. This introduction tries to present ideas in a general way, but is biased towards my work, which targets mostly condensed-matter applications.

In the rest of this chapter, we will introduce the general laws governing many-body systems at the quantum level, and their various approximations, in particular Kohn-Sham density functional theory. We then explain the standard methodology for the simulation of crystals, the pseudopotential method in a plane-wave basis set, and discuss the numerical methods used to solve the resulting equations. We focus in this introductory chapter on finite molecular systems; infinite periodic systems (crystals) are discussed in Chapter III.

II.1. Quantum mechanics of many-body systems

Consider a system of N electrons and M nuclei, coupled through the Coulomb interaction. We denote by x_i the coordinates of the electrons, and by R_a those of the nuclei. As is usual in the mathematical theory of such systems, we ignore spin degrees of freedom for simplicity; spin is of course important in practice, but is easily incorporated into the theory. We use atomic units in which the mass and charge of the electron as well as the reduced Planck constant are one.

The quantum state of N electrons and M nuclei is represented by a normalized element of the Hilbert space

$$\mathcal{H} = \left(\otimes_A^N L^2(\mathbb{R}^3, \mathbb{C})\right) \otimes \left(\otimes^M L^2(\mathbb{R}^3, \mathbb{C})\right).$$

where $\otimes_A^N L^2(\mathbb{R}^3, \mathbb{C})$ is the N-fold antisymmetrized tensor product, which can be identified with the set of functions in $L^2(\mathbb{R}^{3N}, \mathbb{C})$ satisfying the antisymmetry condition

$$\psi(x_{\sigma_1}, \ldots, x_{\sigma_N}) = \varepsilon(\sigma)\psi(x_1, \ldots, x_N) \tag{1}$$

for all permutations σ, and where $\varepsilon(\sigma)$ is the signature of the permutation. The elements of $\mathcal{H}$ are functions of the $3N$ electronic and $3M$ nuclear degrees of freedom; their square modulus represents the probability of finding the nuclei and electrons at a given position. We have here considered nuclei to be distinguishable particles, and ignored all relativistic effects. This is a very good approximation for ordinary chemistry, with the notable exception of relativistic effects on core electrons (see [Pyy12] for a review).

The Hamiltonian is

$$H = \sum_{i=1}^{N} -\frac{1}{2}\Delta_{x_i} + \sum_{a=1}^{M} -\frac{1}{2m_a}\Delta_{R_a} - \sum_{i=1}^{N}\sum_{a=1}^{M} \frac{q_a}{|x_i - R_a|} + \sum_{1\le a<b\le M} \frac{q_a q_b}{|R_a - R_b|} + \sum_{1\le i<j\le N} \frac{1}{|x_i - x_j|} \tag{2}$$

where q_a and m_a are the charge and mass of atom a. The terms correspond respectively to the kinetic energy of the electrons and nuclei, and the electron-nuclei, nuclei-nuclei and

electron-electron interactions. The time-dependent Schrödinger equation is

$$i\partial_t \psi = H\psi. \tag{3}$$

It is remarkable that these two simple equations (2) and (3) contain the answers to the questions asked above, and many more besides.

The main theoretical and practical problem is that of entanglement. A quantum state of the many-body Hamiltonian H is an object in (a subspace of) $L^2(\mathbb{R}^{3(N+M)}, \mathbb{C})$, which contains too much information to be useful. One cannot speak of the state of a single particle without specifying the state of all the others, all particles being coupled together through the Coulomb interaction. Accordingly, a direct numerical discretization of the Schrödinger equation has a state space whose dimension grows exponentially with $N+M$, which makes it intractable even for small molecules. As was remarked by Paul Dirac in 1929, "the underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble. It therefore becomes desirable that approximate practical methods of applying quantum mechanics should be developed, which can lead to an explanation of the main features of complex atomic systems without too much computation" [Dir29].

The first simplication arises because the nuclei are much heavier than the electrons (the proton-to-electron mass ratio is about 1836), so that $m_a \gg 1$. Two mathematically distinct phenomena then occur: first, the dynamics of the electrons and nuclei is approximately adiabatically decoupled. Let

$$H_{\mathrm{el}}(R) = \sum_{i=1}^{N} -\frac{1}{2}\Delta_{x_i} - \sum_{i=1}^{N}\sum_{a=1}^{M} \frac{q_a}{|x_i - R_a|} + \sum_{1\le a<b\le M} \frac{q_a q_b}{|R_a - R_b|} + \sum_{1\le i<j\le N} \frac{1}{|x_i - x_j|}$$

be the electronic Hamiltonian for a given nuclei configuration R, acting on the electronic degrees of freedom $\otimes_A^N L^2(\mathbb{R}^3, \mathbb{C})$. If this Hamiltonian has a non-degenerate ground state $\psi_{\mathrm{el},R}(x)$ with energy $E_0(R)$, which is well-separated from the rest of the spectrum, the wave function of the total system can be approximated as $\psi(x, R, t) \approx \psi_{\mathrm{nucl}}(R, t)\psi_{\mathrm{el},R}(R, x)$. The effective nuclear Hamiltonian is then

$$H_{\mathrm{nucl}} = \sum_{a=1}^{M} -\frac{1}{2m_a}\Delta_{x_a} + E_0(R)$$

with $E_0(R)$ the ground state energy of $H_{\mathrm{el}}(R)$. One can perform an additional approximation and treat the nuclei semiclassically as point particles, replacing H_{nucl} by the classical Hamiltonian $\sum_{a=1}^{M} \frac{1}{2m_a} p_a^2 + E_0(R)$ on the phase space $\mathbb{R}^{6M}$. This combined approximation is usually known as the Born-Oppenheimer approximation. We refer to [Teu03] and references therein for a mathematical justification of this approximation. It is generally accurate as long as the ground-state of $H_{\mathrm{el}}(R)$ is non-degenerate. By the von Neumann-Wigner theorem, eigenvalue crossings of real Hamiltonians happen generically when two independent parameters are varied [NW29]. This gives rise to conical intersections that play an important role in several chemical processes [Bae06].

Assuming the Born-Oppenheimer approximation, we are left with the task of computing the ground state energy $E_0(V)$ of operators $H_{\text{el}}(R)$, of the form

$$H_V = \sum_{i=1}^{N} \left(-\frac{1}{2}\Delta_{x_i} + V(x_i) \right) + \sum_{1 \le i < j \le N} \frac{1}{|x_i - x_j|}$$

on $\otimes_A^N L^2(\mathbb{R}^3, \mathbb{C})$, for Coulomb potentials

$$V(x) = \sum_{i=1}^{N} \sum_{a=1}^{M} \frac{q_a}{|x_i - R_a|} + \sum_{1 \le a < b \le M} \frac{q_a q_b}{|R_a - R_b|}.$$

This Hamiltonian is self-adjoint with domain

$$\left\{ \psi \in \otimes_A^N L^2(\mathbb{R}^3, \mathbb{C}),\ \Delta\psi \in L^2(\mathbb{R}^{3N}, \mathbb{C}) \right\}.$$

This was proven in 1951 by Kato, by what is now seen as a routine application of the Kato-Rellich theorem [Kat51]. This ensures the well-posedness of the time-dependent Schrödinger equation, as well as a decomposition of the spectrum into eigenvalues and continuous spectrum. When $M \ge N$, the spectrum of H_V is composed of infinitely many eigenvalues E_i, accumulating from below at a limit Σ, and continuous spectrum $[\Sigma, +\infty)$ [RS72].

Provided that $E_0(R)$ can be computed, many quantities of interest can be accessed. For example, minimizing $E_0(R)$ with respect to R yields the equilibrium geometry of a molecule. The vibration frequencies can be obtained by diagonalizing the Hessian $\nabla^2 E_0(R)$ at the equilibrium geometry. More complicated properties can be accessed by perturbing the Hamiltonian (for instance, polarization effects can be examined by adding an electric field). It is therefore of fundamental importance to be able to compute ground states of electronic Hamiltonians.

II.2. Approximation methods

It is instructive to consider the case where the electron-electron interaction is neglected. In this case,

$$\widetilde{H}_V = \sum_{i=1}^{N} \left(-\frac{1}{2}\Delta_{x_i} + V(x_i) \right)$$

is the sum of N copies of the single-body operator $h = -\frac{1}{2}\Delta + V$, acting on each electron separately. When V is a Coulomb operator, the spectrum of h is composed of infinitely many negative eigenvalues $\lambda_1 \le \lambda_2 \le \dots$ accumulating at 0, and positive continuous spectrum. Let $\phi_1, \phi_2, \dots$ be orthonormal associated eigenvectors. Assuming that $\lambda_N < \lambda_{N+1}$, the unique ground state of H is given by the Slater determinant

$$\Psi(x_1, \dots, x_N) = \frac{1}{\sqrt{N!}} \det\left((\phi_i(x_j))_{i,j=1,\dots,N} \right)$$

and the associated eigenvalue is $E = \sum_{i=1}^{N} \lambda_i$. The form of Ψ reflects the *Aufbau* principle of chemistry: the ground state is built by filling the available states in order of energy. This is imposed by the antisymmetry of the wavefunction, which implies the Pauli exclusion principle: the N electrons cannot simultaneously occupy the same electronic state ϕ_1, and so must fill up higher energy states $\phi_1, \dots, \phi_N$.

The study of the single-body operator $-\frac{1}{2}\Delta + V$ is therefore a fundamental starting point to undertand the much more complicated interacting case. The spectral theory of single- and many-body Schrödinger operators, and its relationship with static and dynamic properties, has been a major topic in mathematical physics, and detailed studies have investigated properties such as scattering, resonances or semiclassical asymptotics (see [HS12] for an elementary introduction, and [Sim00] for a review).

In the general case where the Coulomb interaction cannot be neglected, approximation techniques have to be designed. These fall into three broad classes:

- Wavefunction methods, which postulate a specific functional form for the wave function ψ and optimize its parameters to minimize the energy. The simplest method in this class, the Hartree-Fock method, postulates that Ψ is a Slater determinant, as in the non-interacting case, and then optimizes the orbitals ϕ_i to minimize the many-body energy. More sophisticated methods include configuration interaction, coupled cluster and density matrix renormalization group methods. We refer to [HJO14] for a comprehensive overview.
- Quantum Monte Carlo, which uses the Monte-Carlo method to bypass the curse of dimensionality. In its basic form, the simplest representative of that class of method seeks to find the minimum eigenstate of a Hamiltonian H by solving the "imaginary-time" equation

$$\partial_t \psi = -(H - E_0)\psi \tag{4}$$

  where E_0 is an estimate of the ground-state energy. This is nothing but a gradient descent algorithm on the Rayleigh quotient, and has the effect of damping the modes associated with eigenvalues greater than E_0 (excited states). This equation is then solved by exploiting the Feynman-Kac formula to link the high-dimensional partial differential equation (4) with a stochastic differential equation. We refer to [HLR94] for details, and to [CJL06] for a mathematical analysis.
- Density Functional Theory (DFT), which takes as its central object the density

$$\rho(x) = N \int_{\mathbb{R}^{3(N-1)}} |\psi(x, x_2, \dots, x_N)|^2 dx_2 \dots dx_N$$

  and tries to express the energy as a function of ρ only. The most widely used version is Kohn-Sham density functional theory, which uses a fictitious system of independent particles to approximate the energy of a given density [KS65].

All of these methods are, in their basic formulations, completely *ab initio*: they do not involve free parameters that have to be fitted to experimental data. Roughly speaking, quantum Monte Carlo and wavefunction methods can give accurate results for small systems (tens of atoms); modern density functional theory methods scale up to thousands of atoms, but are less systematically improvable. Nevertheless, they have been found to give very good results for many systems of interest, especially in solid-state physics.

In the rest of this document, we will focus on density functional theory.

II.3. Density functional theory

Density functional theory refers to a number of methods that reformulate the task of finding the ground state of the many-body Hamiltonian H_V as a variational principle on the density ρ. They can be ordered as a hierarchy of methods that vary in mathematical

and numerical complexity, from the simple Thomas-Fermi model to sophisticated hybrid and meta-GGA models. We describe briefly this "Jacob's ladder" [BW13].

II.3.1. Density functional theory and the Thomas-Fermi model. We first recall the variational principle for a Hamiltonian

$$H_V = \sum_{i=1}^{N} \left(-\frac{1}{2}\Delta_{x_i} + V(x_i) \right) + \sum_{1 \le i < j \le N} \frac{1}{|x_i - x_j|}$$

with ground state energy E_0:

$$E_0 = \inf_{\psi \in \mathcal{W}} \langle \psi, H_V \psi \rangle,$$

with

$$\mathcal{W} = \left\{ \psi \in \otimes_A^N L^2(\mathbb{R}^3, \mathbb{C}),\ \|\psi\|_{L^2} = 1, \|\nabla \psi\|_{L^2} < \infty \right\}.$$

Since

$$\langle \psi, H_V \psi \rangle = \langle \psi, H_0 \psi \rangle + \int \rho_\psi V,$$

we have

$$E_0 = \inf_{\rho \in X} \left(\int \rho V + F(\rho) \right). \tag{5}$$

with

$$X = \{ \rho_\psi, \psi \in \mathcal{W} \}$$
$$F(\rho) = \inf_{\psi \in \mathcal{W}, \rho_\psi = \rho} \langle \psi, H_0 \psi \rangle$$

The variational principle (5) was first proposed by Hohenberg and Kohn [HK64], then justified under the constrained search formalism above by Levy and Lieb [Lev79; Lie83]. Formally, we have formulated the problem of computing E_0 as a minimization of a functional of the density only. This is a drastic reduction in complexity, as the density is a function of three position variables only. We first note that the set X can be characterized explicitly (the N-representability problem for densities) [Lie83, Theorem 1.2] :

$$X = \left\{ \rho \in L^1(\mathbb{R}^3), \rho \ge 0, \int \rho = N, \sqrt{\rho} \in H^1(\mathbb{R}^3) \right\}.$$

If $N = 1$, this is easy, as for a given ρ one can simply choose $\psi = \sqrt{\rho}$. For $N > 1$, the proof proceeds by constructing a Slater determinant whose orbitals all have the same density $\sqrt{\rho/N}$, but with a modulating phase factor to make them orthogonal to each other.

Of course, computing the functional F explicitly is as hard as solving the original problem. However, F is amenable to approximations. The first such approximation is the Thomas-Fermi model

$$F^{\mathrm{TF}}(\rho) = C_{\mathrm{TF}} \int_{\mathbb{R}^3} \rho^{5/3} + \frac{1}{2} \int_{\mathbb{R}^6} \frac{\rho(x)\rho(x')}{|x - x'|} dx dx'$$

where C_{TF} is the Thomas-Fermi constant [LS77b]. The first term is the Thomas-Fermi kinetic energy $K^{\mathrm{TF}}(\rho)$, and the second the Hartree classical electrostatic energy of a charge density ρ. The rationale for the first term is that $C_{\mathrm{TF}}\rho^{5/3}$ is the kinetic energy per unit volume of the free electron gas (the system of non-interacting electrons with no external

potential) at uniform density ρ. This model is very crude from a chemical point of view: for instance, it does not include atomic shell information. It is in fact a theorem in Thomas-Fermi theory (Teller's theorem, [LS77b, Theorem V.1]) that the energy of a molecule is always greater than the energy of its constituant atoms, which does not allow for chemical bonding. Nevertheless, this model is very useful for a number of reasons:

- It is very simple, both mathematically and numerically;
- It displays some characteristics of real systems, like the screening of electrical charges;
- It is exact for molecules in the limit of infinite atomic charge [LS77b];
- It can be improved by the addition of other terms, forming the basis of *orbital-free DFT*.

II.3.2. The reduced Hartree-Fock model. Instead of the kinetic energy of the free electron gas, one can use the minimal kinetic energy of a system of independent electrons at density ρ. There are two flavors of this, differing in their treatment of degenerate systems. The "standard" scheme takes for the kinetic energy of the density ρ

$$K^{\mathrm{S}}(\rho) = \inf_{\phi_1,\dots,\phi_N} \frac{1}{2}\sum_{i=1}^{N}\int_{\mathbb{R}^3} |\nabla\phi_i|^2 \tag{6}$$

where the infimum is taken over all systems of orbitals ϕ_i in $H^1(\mathbb{R}^3)$ satisfying $\langle\phi_i,\phi_j\rangle_{L^2(\mathbb{R}^3)} = \delta_{ij}$ and $\sum_{i=1}^N |\phi_i|^2 = \rho$. The "extended" scheme [DG12] is most conveniently formulated in terms of one-body density matrices:

$$K^{\mathrm{E}}(\rho) = \inf_{\gamma,\,\rho_\gamma=\rho} \mathrm{Tr}\left(\left(-\frac{1}{2}\Delta\right)\gamma\right). \tag{7}$$

where the infimum is taken over all density matrices γ, bounded self-adjoint operators on $L^2(\mathbb{R}^3)$ satisfying $0 \le \gamma \le 1$, such that $\mathrm{Tr}\,\gamma = N$, $\mathrm{Tr}(|\nabla|\gamma|\nabla|) < \infty$. To each such density matrix γ one can associate a density ρ_γ defined by $\rho_\gamma(x) = \gamma(x,x)$, where $\gamma(\cdot,\cdot)$ is the integral kernel of γ.

The second energy is always lower than the first: $K^{\mathrm{E}}(\rho) \le K^{\mathrm{S}}(\rho)$. To see this, note that for any feasible system of orbitals $(\phi_i)_{i=1,\dots,N}$, one can associate a pure state density matrix by $\gamma_\phi = \sum_{i=1}^N |\phi_i\rangle\langle\phi_i|$. This density matrix satisfies the constraints above, $\rho_{\gamma_\phi} = \sum_{i=1}^N |\phi_i|^2 = \rho$, and $\mathrm{Tr}\left(\left(-\frac{1}{2}\Delta\right)\gamma_\phi\right) = \frac{1}{2}\sum_{i=1}^N \int_{\mathbb{R}^3} |\nabla\phi_i|^2$. However, the density matrix formalism also allows for arbitrary mixed states of the form $\gamma = \sum_{i=1}^\infty f_i|\phi_i\rangle\langle\phi_i|$, with fractional occupation numbers $0 \le f_i \le 1$. Physically, this allows for thermodynamic ensembles of electrons. Mathematically, it has the main advantage of being a convex minimization problem. We will in the rest of this section use this extended formalism. However, in practice both are equivalent in non-degenerate cases, and the "standard" scheme is often more computationally convenient.

The choice (7) for the kinetic energy together with the classical Hartree potential energy yields the reduced Hartree-Fock (rHF) model

$$F^{\mathrm{H}}(\rho) = \inf_{0\le\gamma=\gamma^*\le 1,\,\rho_\gamma=\rho} \mathrm{Tr}\left(\left(-\frac{1}{2}\Delta\right)\gamma\right) + \frac{1}{2}\int_{\mathbb{R}^6} \frac{\rho(x)\rho(x')}{|x-x'|}dxdx'$$

This model is of a form similar to the Hartree-Fock model, without an exchange term. It is also called the Hartree model. It reproduces qualitatively a large number of static

and dynamical phenomena: shell structure, bonding, screening, Friedel oscillations, etc. However, it remains a very crude approximation. The very simplest example of failure is the hydrogen atom, with $N = 1$, where the rHF energy

$$F^{\mathrm{H}}(\rho) = \frac{1}{2}\int_{\mathbb{R}^3} |\nabla\phi|^2 + \frac{1}{2}\int_{\mathbb{R}^6} \frac{\rho(x)\rho(x')}{|x-x'|}dxdx'$$

includes a non-physical self-interaction term. On more complex systems, this leads to wrong predictions: for instance, the H^- ion (two electrons, one proton) is predicted not to be stable (meaning that its rHF energy is higher than that of the atom H and a free electron), contradicting both experiments and more careful computations. Nevertheless, the rHF is a mathematically useful object: it shares many of the properties of the widely used Kohn-Sham model, while taking the form of a convex optimization problem.

II.3.3. Kohn-Sham DFT. The Kohn-Sham method improves on the rHF model by adding an extra term

$$F^{\mathrm{KS}}(\rho) = \inf_{0\leq\gamma=\gamma^*\leq 1,\, \rho_\gamma=\rho} \mathrm{Tr}\left(\left(-\frac{1}{2}\Delta\right)\gamma\right) + \frac{1}{2}\int_{\mathbb{R}^6} \frac{\rho(x)\rho(x')}{|x-x'|}dxdx' + E_{\mathrm{xc}}(\rho).$$

This term $E_{\mathrm{xc}}(\rho)$ is called the *exchange-correlation* energy (recall that the correlation energy is defined as the difference between the exact and Hartree-Fock energy). The first proposed E_{xc} [KS65] is the local density approximation (LDA). It is of the form

$$E_{\mathrm{xc,LDA}}(\rho) = \int_{\mathbb{R}^3} e_{\mathrm{xc,LDA}}(\rho(x))dx$$

and is fitted to exact asymptotics and quantum Monte-Carlo computations to be exact on the uniform electron gas (the interacting electron gas at constant density ρ) [PW92]. We refer to [LLS19] for a recent mathematical proof of the validity of the LDA for almost-uniform systems.

Despite being a crude approximation, the LDA gives impressive results, especially in the solid phase: for instance, it is able to predict the structure and lattice parameters of many crystals to an accuracy of a few percents. However, it severely overestimates the binding energy (well depth) of molecules, and is therefore largely inapplicable in chemistry. This was partially remedied by generalized gradient approximations (GGAs), of the form

$$E_{\mathrm{xc,GGA}}(\rho) = \int_{\mathbb{R}^3} e_{\mathrm{xc,GGA}}(\rho(x), |\nabla\rho(x)|)dx$$

where $e_{\mathrm{xc,GGA}}$ is fitted to reproduce a number of "exact conditions" (behavior in particular limits) [PBE96]. Even more sophisticated methods include the hybrid methods, that incorporate a fraction of the exchange energy of Hartree-Fock theory, and the "meta-GGAs" that include additional information such as the kinetic energy density. Various parametrizations are used, ranging from the purely ab initio, where only fundamental constants are used [BEP97], to functionals with a large number of free parameters that are fit to experimental data [Bec93].

Examples of accuracy of these approximations are given in Figure 1. The accuracy of local and semi-local approximations such as the LDA and GGA are often remarkably accurate for equilibrium properties of solids. The limits of Kohn-Sham density functional theory can be classified as follows:

	LDA	GGA	Hybrid
Bond length	0.3%	0.4%	0.8%
Binding energy	18%	7%	0.8%

TABLE 1. Accuracy relative to experiment of bond length ℓ and binding energy E_{binding} for diazote, N_2. If $E(L)$ is the energy of two azote atoms at a distance L, then $\ell = \arg\min E(R)$, and $E_{\text{binding}} = E(\infty) - E(\ell)$. Data from [BW13].

- Failure to reproduce dynamical properties and excited states. This is intrinsic to DFT, which is a ground state theory. Theories based on many-body perturbation theory such as the GW method [AG98] or on dynamics such as time-dependent DFT (TDDFT) [BWG05] allow for the computation of properties such as excitation energies or absorption spectra. They are often based on a preliminary DFT ground state computation.
- Failure to predict the structure of strongly correlated materials. For instance, Mott insulators such as NiO are predicted by any mean-field theories such as Kohn-Sham density functional theory to be metals (because they have an odd number of electrons per unit cell and so necessarily have unfilled bands), but are experimentally found to be insulators. Theories explicitly incorporating many-body effects such as the Dynamical Mean-Field Theory (DMFT) [Geo04] can be combined with DFT to explain the behavior of such systems.
- Failure to reproduce many-body effects, such as dispersion forces, bond dissociation, or superconductivity. Semi-empirical models can be used to incorporate some effects such as dispersion into DFT [Gri11].

The models presented above can be ordered in the following way. First, Thomas-Fermi type models provide some qualitative properties of the electron gas, but do not reproduce chemistry at all. The rHF model explains bonding and shell structure, but is quantitatively wrong. Simple purely ab initio models such as LDA DFT or Hartree-Fock are quantitatively correct for some static properties, but not for others. Advanced DFT (using empirical corrections) can be accurate for static properties of most materials, but fails on strongly correlated systems. More complex electronic structure methods, such as wavefunction methods, quantum Monte-Carlo, many-body perturbation theory or renormalization group approaches, are often in good agreement with experiment, but their applicability remains limited to small systems. Kohn-Sham DFT occupies a "sweet spot" of reasonable accuracy for a relatively modest computational cost, and is therefore the workhorse of condensed matter physics.

From the mathematical point of view, these models are nonlinear, in contrast to the original linear many-body Schrödinger. This is a source of major complications: the mathematical apparatus switches from spectral theory to the heavier calculus of variations, and uniqueness is lost in non-convex models. At the same time, the relative simplicity of the models (compared to the full many-body Schrödinger equation) means that a wider range of qualitative and asymptotic properties of solutions is accessible to analysis. The Thomas-Fermi model has been comprehensively studied from the '70s on, with a very complete description of screening, (lack of) binding, and derivations in various asymptotic regimes

[LS77b]. Extensions to variations such as the Thomas-Fermi-von Weizsäcker model soon followed [BBL81]. The more complex Hartree and Hartree-Fock models were studied in [LS77a; Lio87], where existence and qualitative properties were given. The LDA and GGA KSDFT models were studied in [LB93; AC09].

In the following, we assume that a particular flavour of KSDFT (for instance LDA or GGA) has been selected, and consider the model

$$\inf_{0\leq\gamma=\gamma^*\leq 1,\,\mathrm{Tr}\,\gamma=N} E(\gamma) \tag{8}$$

with

$$E(\gamma) = \mathrm{Tr}\left(\left(-\frac{1}{2}\Delta + V\right)\gamma\right) + \frac{1}{2}\int_{\mathbb{R}^6} \frac{\rho_\gamma(x)\rho_\gamma(x')}{|x-x'|}dxdx' + E_{\mathrm{xc}}(\rho_\gamma).$$

II.4. The Kohn-Sham equations

Consider a minimizer γ_0 of (8). Since the minimization set is convex, for any $\widetilde{\gamma}$ such that $0 \leq \widetilde{\gamma} = \widetilde{\gamma}^* \leq 1$, $\mathrm{Tr}\,\widetilde{\gamma} = N$ and $t \in [0,1]$, we have $E(\gamma_0 + t(\widetilde{\gamma} - \gamma_0)) \geq E(\gamma_0)$. Differentiating this at $t = 0$, we obtain

$$\gamma_0 \in \operatorname*{arg\,min}_{0\leq\gamma=\gamma^*\leq 1,\,\mathrm{Tr}\,\gamma=N} \mathrm{Tr}(H_{\gamma_0}\gamma) \tag{9}$$

where the self-consistent Hamiltonian H_γ (also called *Fock matrix* in quantum chemistry) is

$$H_\gamma = -\frac{1}{2}\Delta + V + \rho_\gamma * \frac{1}{|x|} + V_{\mathrm{xc}}(\rho_\gamma) \tag{10}$$

where V_{xc} is the gradient of E_{xc}.

It is easily seen that

$$\operatorname*{arg\,min}_{0\leq\gamma=\gamma^*\leq 1,\,\mathrm{Tr}\,\gamma=N} \mathrm{Tr}(H_{\gamma_0}\gamma) = \mathbb{1}(H_{\gamma_0} < \mu) + \delta$$

where δ is in the spectral subspace associated with μ of H_{γ_0}. In particular, if the bottom of the spectrum of H_{γ_0} is composed of $N+1$ eigenvalues $\lambda_1, \ldots, \lambda_{N+1}$ with $\lambda_{N+1} > \lambda_N$, then

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

where ϕ_i are orthonormal eigenvectors associated with $\lambda_1, \ldots, \lambda_N$. This is reminiscent of the situation for a non-interacting system described in Section II.2: the electrons fill the energy levels in order, according to the Aufbau principle.

We will say that a system has the Aufbau property if all the minimizers γ of (8) are such that H_γ is gapped, in the sense that $\lambda_{N+1} > \lambda_N$. If the system has the Aufbau property, then the extended scheme (8) is equivalent to

$$\inf_{\langle\phi_i,\phi_j\rangle=\delta_{ij}} E(\phi_1, \ldots, \phi_N), \tag{11}$$

with

$$E(\phi_1, \dots, \phi_N) = \int_{\mathbb{R}^3} \left(\sum_{i=1}^{N} \frac{1}{2} |\nabla \phi_i|^2 + V \rho_\phi \right) + \frac{1}{2} \int_{\mathbb{R}^6} \frac{\rho_\phi(x) \rho_\phi(x')}{|x - x'|} dx dx' + E_{\mathrm{xc}}(\rho_\phi),$$

$$\rho_\phi = \sum_{i=1}^{N} |\phi_i|^2.$$

When the Aufbau property is false, both models are not equivalent, and (8) is a strict relaxation of (11).

In the strictest sense, for a given exchange-correlation functional, there are no particular reason to prefer (8) to (11): both are equally wrong. In practice, many systems are found to be gapped, and this is not an issue. From the numerical point of view, both models are equally problematic when degeneracy is present; this is often remedied with *smearing*, an artificial small temperature (see Section III.2.3). Mathematically however, it is more convenient to work with (8) than (11), because the minimization set is then convex. This gets rid of one potential source of multiple minimizers, although it still does not result in a fully convex problem because E_{xc} is not convex. In the rest of this document, we will focus for simplicity on the standard Kohn-Sham model (11), and assume the Aufbau property.

The Euler-Lagrange equations for the problem (11) are

$$H_\Phi \phi_i = \sum_{j=1}^{N} \lambda_{ij} \phi_j,$$

where λ_{ij} is the Lagrange multiplier associated to the constraint $\langle \phi_i, \phi_j \rangle = \delta_{ij}$, and where $H_\Phi = H_{\gamma_\Phi}$ with $\gamma_\Phi = \sum_{i=1}^{N} |\phi_i\rangle\langle\phi_i|$ the density matrix associated with Φ. We can simplify this equation by taking advantage of the orthogonal invariance of the energy: $E(\Phi) = E(\gamma_\Phi)$ only depends on the density matrix γ_Φ, which itself depends on the subspace spanned by the ϕ_i, and not the choice of basis in this subspace. Accordingly, for every $N \times N$ unitary matrix U, we have $E(\Phi U) = E(\Phi)$. We can exploit this to diagonalize the Hermitian matrix λ_{ij}, and obtain, up to rotation of the ϕ_i,

$$H_\Phi \phi_i = \lambda_i \phi_i. \tag{12}$$

Furthermore, assuming the Aufbau property, the λ_i are the lowest eigenvalues of H_Φ. This is physically intuitive because it matches the case of a system of N non-interacting fermions, where the many-body ground state is given by a Slater determinant built using the first N eigenstates of the single-particle Hamiltonian. In the Kohn-Sham model however, the single-particle Hamiltonian H_Φ depends self-consistently on the orbitals ϕ_i. It can therefore be seen as a *mean-field* theory of the type found in statistical mechanics [CL00]: the electrons behave as independent particles in the mean-field potential they self-consistently create.

This interpretation as a mean-field theory is also conceptually important, because it provides a justification for the chemical picture of orbitals and the shell structure of atoms, which are lost in the purely many-body formalism. By combining exact computations on the hydrogen atom with heuristic arguments on screening through the mean-field operator, it explains the rules of chemistry, such as the Madelung "$n + \ell$" ordering rule, or Hund's rule that electrons tend to occupy degenerate orbitals in parallel spins.

II.5. Plane-wave discretization

We now consider the discretization of (11). A wide array of methods are used in practice: localized basis sets, finite differences, finite elements, wavelets, spectral methods, etc. We focus here on the plane-wave method, the most widely used method in condensed matter applications. Although this method was originally designed for computing properties of crystals and is still mostly used for that purpose, it can also be used for molecules. For pedagogical purposes we introduce it here in the case of an isolated molecule. The extension to infinite periodic systems (crystals) is presented in Chapter III.

Note that, since for an isolated system the self-consistent potential tends to zero at infinity, the solutions ϕ_i of the self-consistent equation (12) for negative λ_i are exponentially localized (with characteristic length $1/\sqrt{-\lambda_i}$). It is therefore justified, with exponentially small error, to limit ourselves to a large box, which we will take for simplicity to be $\Gamma = \left[-\frac{L}{2}, \frac{L}{2}\right]^3$. We will replace the space $\mathbb{R}^3$ by Γ equipped with the topology of a torus (therefore imposing periodic boundary conditions). We can then expand orbitals ϕ_i in the orthonormal Fourier basis $\{e_K\}_{K\in\mathcal{R}_L^*}$, with

$$e_K(x) = \frac{1}{\sqrt{|\Gamma|}} e^{iK\cdot x}$$

and $\mathcal{R}_L^* = \frac{2\pi}{L}\mathbb{Z}^3$. We then have the expansion

$$\phi_i(x) = \sum_{K\in\mathcal{R}_L^*} c_{iK} e_K(x).$$

The kinetic, potential and exchange-correlation terms in (11) adapt easily by simply truncating the integrals to Γ. To approximate the Hartree term in a way that leads to simple computations, we replace the Coulomb kernel $\frac{1}{|x|}$ by the periodic Coulomb kernel [LS77b]

$$G_L(x) = \sum_{K\in\mathcal{R}_L^*\setminus\{0\}} \frac{4\pi e_K(x)}{|K|^2}.$$

Then, we seek to minimize the total energy

$$E_L(\Phi) = \int_\Gamma \left(\frac{1}{2}\sum_{i=1}^N |\nabla\phi_i|^2 + V\rho_\phi\right) + \frac{1}{2}\int_{\Gamma\times\Gamma} \rho_\phi(x)\rho_\phi(x')G_L(x,x')dxdx' + E_{\rm xc}(\rho_\phi)$$

variationally by limiting the discretization space $\{e_K\}_{K\in\mathcal{R}_L^*}$ to the subspace

$$\mathcal{X} = \mathrm{Span}\left(e_K, K\in\mathcal{R}_{E_{\rm cut}}^*\right), \qquad \mathcal{R}_{E_{\rm cut}}^* = \left\{K\in\mathcal{R}_L^*, \frac{1}{2}|K|^2 \le E_{\rm cut}\right\}$$

where $E_{\rm cut} > 0$ is a truncation parameter that controls the maximum kinetic energy allowed in the system.

The resulting energy as a function of the coefficients c_{iK} is

$$\begin{aligned} E_L(c) = &\sum_{\substack{i=1,\dots,N \\ K,K' \in \mathcal{R}^*_{E_{\text{cut}}}}} \left(\frac{1}{2}|K|^2 \delta_{KK'} + \widehat{V}_{K-K'} \right) \overline{c_{iK}} c_{iK'} \\ &+ \sum_{\substack{i,j=1,\dots,N \\ K,K' \in \mathcal{R}^*_{E_{\text{cut}}} \\ K \neq K'}} \frac{4\pi \overline{c_{iK}} c_{iK'} \overline{c_{jK}} c_{jK'}}{|K'-K|^2} \\ &+ E_{\text{xc}} \left(\sum_{\substack{i=1,\dots,N \\ K,K \in \mathcal{R}^*_{E_{\text{cut}}}}} \overline{c_{iK}} c_{iK'} e_{K'-K}(x) \right) \end{aligned}$$

where

$$\widehat{V}_K = \frac{1}{\sqrt{|\Gamma|}} \int_\Gamma e^{-iKx} V(x) dx$$

are the Fourier coefficients of the periodic extension of V.

If the formulas above are evaluated naively, the computation time scales quadratically with both N and the number of plane waves $|\mathcal{R}^*_{E_{\text{cut}}}|$. However, these expressions are convolutions, arising from pointwise multiplication in real space. These convolutions can be evaluated efficiently in real space, using the discrete convolution theorem: the cyclic convolution of two arrays can be computed by a discrete Fourier transform, evaluated efficiently using fast Fourier transforms (FFT). To avoid aliasing effects arising from the cyclic convolution, zero-padding is used: the discrete Fourier transforms are performed on a Cartesian grid that contains all the $K + K'$, for $K, K' \in \mathcal{R}^*_{E_{\text{cut}}}$.

It is important to note that this procedure allows us to treat the kinetic, potential and Hartree terms *exactly*: for any given set of coefficients c_{iK}, the energy terms computed in this way are the exact energy terms of the orbitals $\phi_i(x) = \sum_{K \in \mathcal{R}^*_{E_{\text{cut}}}} c_{iK} e_K(x)$. Therefore, for the rHF model, the plane-wave method (for a given L) is *variational*: the ground state energy decreases with E_{cut}. However, the exchange-correlation term is a non-polynomial function of the density, and cannot be evaluated exactly in this fashion. In practice, this term is approximated by an integration on the same real-space grid as for the other terms.

The numerical analysis of this method was performed in [CCM12] (see [DM17] for a-posteriori error bounds on the related Gross-Pitaevskii equation). Note that this method is based on an expansion of the orbitals ϕ_i in a Fourier basis. In order for this to be effective, the orbitals need to be smooth, because of the equivalence between smoothness in real space and decay in reciprocal space. However, the singularity of the Coulomb potential imposes *cusps* on the ϕ_i: for instance, the first eigenfunction of the Hydrogen atom is proportional to $e^{-|x|}$. Further, even if we remove these cusps by mollifying the Coulomb potential, the ϕ_i still need to oscillate to satisfy the orthogonality conditions $\langle \phi_i, \phi_j \rangle = \delta_{ij}$, and these oscillations need a large number of plane waves to represent properly. The plane-wave discretization is therefore not applicable directly to atomic systems; in practice, these problems are remedied through the use of the pseudopotential approximation.

II.6. Pseudopotentials

The present discussion is focused on the use of pseudopotential method in solid-state physics. We refer to [Gia10] and references therein for an overview of the pseudopotential method in practice. A very similar method is used in quantum chemistry, where it goes by the name of "effective core potential"; we refer to [Dol+00] for details. To simplify the discussion, we assume an isolated system of non-interacting electrons (we ignore the Hartree and exchange-correlation terms). With appropriate modifications, pseudopotentials can be extended to treat the Kohn-Sham equations.

II.6.1. Orbitals and chemical bonds. Reintroducing spin temporarily, recall that the states of an atomic radial Hamiltonian $H_{\text{at}} = -\frac{1}{2}\Delta + V_{\text{at}}$ can be labelled as $\phi_{n\ell m\sigma}$:

$$H_{\text{at}}\phi_{n\ell m\sigma} = \varepsilon_{n\ell}\phi_{n\ell m\sigma}$$
$$\phi_{n\ell m\sigma}(x) = \frac{R_{n\ell}(|x|)}{|x|} Y_{\ell m}\left(\frac{x}{|x|}\right)\sigma,$$

where $n \geq 1$, $\ell \geq 0$, $m = -\ell, \dots, \ell$ and $\sigma = \{\alpha, \beta\}$ are the principal, azimuthal, magnetic and spin quantum numbers. The functions $Y_{\ell m}$ are the real spherical harmonics. The functions $R_{n\ell}$ are solutions of the radial Schrödinger equation

$$-\frac{1}{2}R''_{n\ell}(r) + \left(\frac{\ell(\ell+1)}{2r^2} + V_{\text{at}}(r)\right) R_{n\ell}(r) = \varepsilon_{n\ell} R_{n\ell}(r)$$

for $r > 0$, with $R_{n\ell}(0) = 0$ and $\int_0^\infty |R_{n\ell}|^2 = 1$.

There are $2(2\ell+1)$ available states $(\phi_{n\ell m\sigma})_{m=-\ell,\dots,\ell,\sigma=\{\alpha,\beta\}}$ for a given energy level $\varepsilon_{n\ell}$. The electronic state of an atom is conventionally given as a sequence of terms of the form $(n+\ell)\ell^k$, meaning that k of the $4\ell+2$ available states with a given (n, ℓ) are occupied. The angular momentum quantum number ℓ is labelled using letters: s, p, d, f, etc. For instance, Silicon has 14 electrons, and its electronic structure is $1s^2\ 2s^2 2p^6\ 3s^2 3p^2$. This means: two electrons in the $\phi_{n=1,\ell=0,m=0,\sigma=\{\alpha,\beta\}}$ orbitals, two electrons in the $\phi_{n=2,\ell=0,m=0,\sigma=\{\alpha,\beta\}}$ orbitals, six electrons in the $\phi_{n=1,\ell=1,m=\{-1,0,1\},\sigma=\{\alpha,\beta\}}$ orbitals, etc. The spatial extension $\langle r\rangle = \int_{\mathbb{R}^3} |x||\phi_i|^2(x)dx$ and energies ε of these states, computed using the PBE density functional [PBE96], are given Table 2, and the orbitals are plotted Figure 1 (left panel).

State	(n, ℓ)	$\varepsilon_{n\ell}$ (Ha)	$\langle r\rangle$ (Bohr)
$1s^2$	$(1, 0)$	-64.4	0.11
$2s^2$	$(2, 0)$	-5.10	0.57
$2p^6$	$(1, 1)$	-3.51	0.54
$3s^2$	$(3, 0)$	-0.40	2.17
$3p^2$	$(2, 1)$	-0.15	2.79

TABLE 2. Energy ε and spatial extension $\langle r\rangle$ for the orbitals of the Silicon atom at the PBE level. Data from the `atomic` code included in the Quantum Espresso distribution. For comparison, the energy of a typical covalent bond is about 0.1 Ha, and the interatomic distance in bulk silicon at ambient conditions is about 4.5 Bohr.

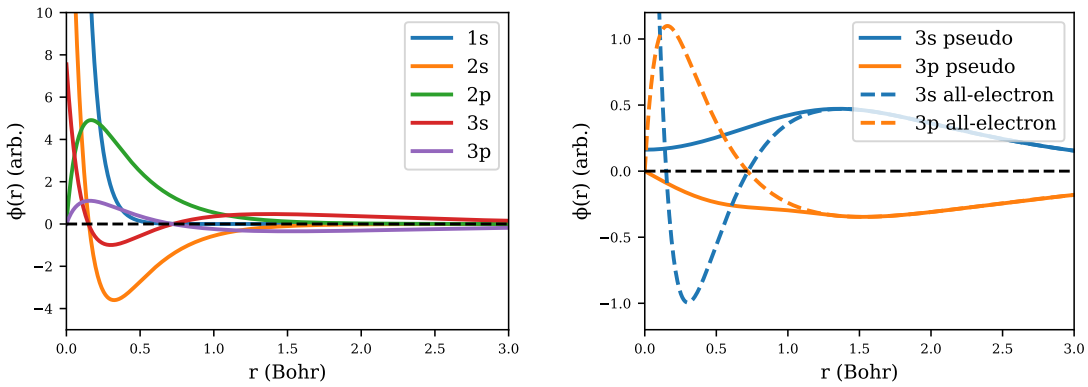


FIGURE 1. Orbitals of the Si atom at the PBE level (left), and pseudized $3s$ and $3p$ orbitals (right). The pseudized orbitals are nodeless, and free of cusps.

From this data, it can be seen that electrons can be clearly separated in two classes. The 10 electrons in the $1s$, $2s$ and $2p$ configurations are *core* electrons. They are mainly localized close to the nucleus, with a low energy. This means that they react very little to their environment, as can be shown using the following heuristic argument. Assume that atom A has only one orbital ϕ_A, with energy E_A, and that atom B has only one orbital ϕ_B, with energy $E_B > E_A$. When in the diatomic configuration, the total Hamiltonian, expanded in the basis of the ϕ_A and ϕ_B, has the approximate form

$$\begin{pmatrix} E_A & \alpha \\ \alpha & E_B \end{pmatrix}$$

where the coupling term α is due to the overlap between ϕ_A and ϕ_B. The eigenvalues of this matrix are

$$E_\pm = \frac{E_A + E_B}{2} \pm \sqrt{\left(\frac{E_B - E_A}{2}\right)^2 + \alpha^2}.$$

In order to have $E_\pm$ significantly different from the isolated energies E_A and E_B (which leads to chemical bonding), we need α to be of comparable magnitude to $E_B - E_A$. In other words, to get a bond between two atoms, there need to be orbitals with a significant overlap which are close in energy.

To a very good approximation, we can therefore consider the $1s, 2s$ and $2p$ orbitals of Silicon as *core* orbitals, that are frozen: they are the same in any chemical environment. By contrast, the $3s$ and $3p$ orbitals are *valence* orbitals, and bind to the orbitals of other atoms.

It follows from this splitting that representing all the ϕ_i is wasteful, since the core electrons do not contribute significantly to chemical bonding. On the other hand, the valence electrons still feel the influence of the core electrons, and have strong oscillations near the nuclei. Combined with the cusp at the nucleus, this makes them very hard to represent in a plane wave basis. A further complication is that the core electrons of heavy atoms, having a high velocity, are subject to non-negligible relativistic effects. This

impacts the valence orbitals indirectly through the orthogonalization requirements and the mean-field potential. All these difficulties are remedied through pseudopotentials, at the cost of an approximation.

II.6.2. The pseudopotential method. To introduce pseudopotentials, we first split the energy states $\varepsilon_{n\ell}$ of the atom into core and valence orbitals. We ignore spin in the following discussion, and label the orbitals only by $\phi_{n\ell m}$. We take $\ell = 0, \dots, \ell_{\max}$ to be set the of angular momenta for which there is a valence orbital, $\mathcal{I}_\ell = \{n \in \mathbb{N}, \varepsilon_{n\ell} \text{ is a valence state}\}$ and $n_{\text{core},\ell} = \min \mathcal{I}_\ell - 1$ the number of core orbitals with angular momentum ℓ. For instance, for the Silicon atom a natural choice is to set the $1s, 2s$ and $2p$ orbitals as core, and the $3s$ and $3p$ orbitals as valence: $\ell_{\max} = 1, \mathcal{I}_0 = \{3\}, \mathcal{I}_1 = \{2\}$, $n_{\text{core},0} = 2$, $n_{\text{core},1} = 1$.

The pseudopotential method then replaces the atomic potential V_{at} by a pseudopotential $\widetilde{V}_{\text{at}}$ (usually a nonlocal operator) such that the lowest energy eigenfunctions of the pseudo-Hamiltonian $\widetilde{H}_{\text{at}} = -\frac{1}{2}\Delta + \widetilde{V}_{\text{at}}$, the pseudo-orbitals $\widetilde{\phi}_{n\ell m}$ for $\ell = 0, \dots, \ell_{\max}, n = 1, \dots, |\mathcal{I}_\ell|$, are smooth and match the valence orbitals $\phi_{n+n_{\text{core},\ell},\ell m}$ of the real Hamiltonian $H_{\text{at}} = -\frac{1}{2}\Delta + V_{\text{at}}$ outside the core of the atom. In Silicon, this means for instance that the new $1s$ state of $\widetilde{H}$ matches the $3s$ state of H outside a chosen cutoff radius r_{c}. Since the decay rate of an eigenstate with a negative eigenvalue ε is $\sqrt{-\varepsilon}$, it follows that the eigenvalues of $\widetilde{H}_{\text{at}}$ should match those of H_{at}:

$$\widetilde{H}_{\text{at}} \widetilde{\phi}_{n\ell m} = \varepsilon_{n+n_{\text{core},\ell},\ell}\, \widetilde{\phi}_{n\ell m}$$

for $\ell = 0, \dots, \ell_{\max}, n = 1, \dots, |\mathcal{I}_\ell|, m = -\ell, \dots, \ell$.

Once this procedure is established for an atom, all the individual potentials of the atoms in a molecule can be replaced ("pseudized") in this fashion, and the pseudo-Hamiltonian of the molecule is then solved.

Setting aside the construction of these pseudopotentials for a moment, two main issues have to be considered. First, *transferability*: to what extent do the results of a pseudopotential computation on a molecule match the results of the corresponding reference (all-electron) computation? Second, *softness*: how many Fourier modes are needed to represent the pseudo-orbitals accurately? Pseudopotential seek an optimal compromise between these two competing objectives. Transferability is usually assessed empirically, by comparing the properties of a molecular system to those obtained using an all-electron computation. It is improved by choosing a small cutoff radius r_{c} and ensuring that all relevant electrons are included as valence. Softness is ensured by choosing a smooth pseudopotential $\widetilde{V}_{\text{at}}$ and by the fact that the lowest energy eigenfunction of a radial local Hamiltonian is nodeless, which reduces oscillations (see Figure 1).

Building pseudopotentials is often done in two parts: given a local potential $V_{\text{at}}(x) = -\frac{Z}{|x|}$, first construct smooth pseudo-orbitals $\widetilde{\phi}_{n\ell m}$, and second construct a pseudopotential $\widetilde{V}_{\text{at}}$ such that the eigenstates of $\widetilde{H}$ are the $\widetilde{\phi}_{n\ell m}$.

The radial part $\widetilde{R}_{n\ell}$ of the pseudo-orbitals $\widetilde{\phi}_{n\ell m}$ should match the original orbitals outside of a cutoff radius:

$$\widetilde{R}_{n\ell}(x) = R_{n+n_{\text{core},\ell},\ell} \qquad \text{for } |x| \geq r_{\text{c}}.$$

As eigenstates of $\widetilde{H}$, they should also be orthonormal. Orbitals with different angular momenta are automatically orthogonal; the orbitals with the same angular momenta must satisfy

$$\langle \widetilde{R}_{m\ell}, \widetilde{R}_{n\ell} \rangle_{L^2(\mathbb{R}^+)} = \delta_{mn}$$

which implies the "generalized norm-conservation condition"

$$\int_0^{r_c} \overline{\widetilde{R}_{n\ell}(r)} \widetilde{R}_{n'\ell}(r) dr = \int_0^{r_c} \overline{R_{n+n_{\mathrm{core},\ell},\ell}(r)} R_{n'+n_{\mathrm{core},\ell},\ell}(r) dr \tag{13}$$

for all $m, n \in \mathcal{I}_\ell$. Finally, they should be smooth, and, as low-energy eigenfunctions, have the correct amount of nodes.

Historically, most pseudopotentials have been constructed assuming the pseudization of only one orbital per angular momentum: $|\mathcal{I}_\ell| = 1$ for $\ell = 0, \ldots, \ell_{\max}$. This is for instance reasonable in the Silicon example above. For simplicity, we denote these orbitals and pseudo-orbitals by $\phi_{\ell m}$ and $\widetilde{\phi}_{\ell m}$, with energies ε_ℓ. The Troullier-Martins procedure [TM91] then postulates a simple functional form for $\widetilde{R}_\ell$, and adjusts its parameters to match the values and the derivatives up to fourth order of $\widetilde{R}_\ell$ and R_ℓ at r_c. The RRKJ procedure expands $\widetilde{R}_\ell$ in a basis of Bessel functions, and minimizes their kinetic energy above a certain cutoff energy to ensure smoothness [RRKJ90].

Once the $\widetilde{\phi}_{\ell m}$ are constructed, a potential $\widetilde{V}_{\mathrm{at}}$ is constructed. A solution is to first choose the local radial potential $\widetilde{V}_\ell$ so that $\widetilde{\phi}_{\ell m}$ is a solution of

$$\left(-\frac{1}{2}\Delta + \widetilde{V}_\ell\right) \widetilde{\phi}_{\ell m} = \varepsilon_\ell \widetilde{\phi}_{\ell m}.$$

This can be done by simply solving the equation above for $\widetilde{V}_\ell$. Then, we can choose the semilocal form

$$\widetilde{V}_{\mathrm{at}} = \sum_{\ell=0}^{\ell_{\max}} P_\ell \widetilde{V}_\ell P_\ell, \tag{14}$$

where $P_\ell = \sum_{m=-\ell,\ldots,\ell} |Y_{\ell m}\rangle\langle Y_{\ell m}|$ is the projector on states with angular momentum ℓ. This operator automatically has the $\widetilde{\phi}_{\ell m}$ as eigenstates with correct eigenvalues, completing the pseudo-potential.

This procedure however has a major defect: the operator $\widetilde{V}_{\mathrm{at}}$ is a nonlocal operator whose evaluation in plane-wave codes requires the computation of a dense matrix. This problem was solved by Kleinmann and Bylander [KB82], who realized that the above procedure was wasteful; the most economical operator that will produce the $\widetilde{\phi}_{\ell m}$ as eigenstates is the low-rank form

$$\widetilde{V}_{\mathrm{at}}^{\mathrm{KB}} = \widetilde{V}_{\mathrm{loc}} + \sum_{\ell=0}^{\ell_{\max}} \sum_{m=-\ell}^{\ell} \frac{|\beta_{\ell m}\rangle\langle\beta_{\ell m}|}{\langle \beta_{\ell m}, \widetilde{\phi}_{\ell m}\rangle} \tag{15}$$

where $\widetilde{V}_{\mathrm{loc}}$ is an arbitrary radial local potential and

$$\beta_{\ell m} = \left(\varepsilon_\ell - \left(-\frac{1}{2}\Delta + \widetilde{V}_{\mathrm{loc}}\right)\right) \widetilde{\phi}_{\ell m}.$$

By construction, $(-\frac{1}{2}\Delta + \widetilde{V}_{\mathrm{at}}^{\mathrm{KB}})\widetilde{\phi}_{\ell m} = \varepsilon_\ell \widetilde{\phi}_{\ell m}$, as desired. The potential $\widetilde{V}_{\mathrm{loc}}$ is usually chosen to remove one of the ℓ components from the sum.

Note that we have only insisted on the fact that the $\widetilde{\phi}_{\ell m}$ are eigenfunctions of $\widetilde{H}_{\text{at}}$ at energies ε_ℓ, but have said nothing of the other states; it may very well happen that $\widetilde{H}_{\text{at}}$ has states with lower energies ("ghosts"), invalidating the pseudopotential. This is usually treated heuristically, by varying parameters in the pseudopotential construction to avoid the apperance of ghosts.

The procedure outlined above is the basis for the construction of the historically successful Troullier-Martins and RRKJ pseudopotentials. More modern pseudopotentials try to reproduce more than one state per angular momentum. This is especially useful for highly-localized valence orbitals (like the $2p$ orbitals of oxygen), or semicore orbitals, which are not well isolated from the valence orbitals. In this case, one cannot use the semilocal form (14), because no single potential V_ℓ can reproduce all the pseudo-orbitals at the required energies. However, the form (15) naturally generalizes to this case: we seek a low-rank potential such that $\widetilde{H}_{\text{at}}\widetilde{\phi}_{n\ell m} = \varepsilon_{n\ell}\widetilde{\phi}_{n\ell m}$ for $i = 1, \dots, n_\ell$, and obtain

$$\widetilde{V}_{\text{at}}^{\text{KB}} = \widetilde{V}_{\text{loc}} + \sum_{\ell=0}^{\ell_{\max}} \sum_{n=1}^{n_\ell} \sum_{m=-\ell}^{\ell} B_{nn'\ell m} |\beta_{n\ell m}\rangle\langle\beta_{n\ell m}| \tag{16}$$

where the $\beta_{n\ell m}$ are defined as above, and

$$B_{nn'\ell m} = \langle \widetilde{\phi}_{n\ell m}, \beta_{n'\ell m} \rangle.$$

From the generalized norm-conservation property (13), $\langle \widetilde{\phi}_{n\ell m}, \widetilde{\phi}_{n'\ell m} \rangle = \delta_{nn'}$, and it follows that B is Hermitian.

The scheme above was used by Hamann [Ham13], who constructed pseudo-orbitals satisfying the generalized norm-conservation property (13). That work built on an earlier method by Vanderbilt, the "ultra-soft pseudopotentials", which relaxed the constraint (13) by introducing a modified inner product in which (13) holds, resulting in a generalized eigenvalue problem.

An alternative to norm-conserving or ultra-soft pseudopotentials is the projector-augmented wave (PAW) method by Blöchl [Blö94], which uses a transformation that allows the recovery of the original all-electron orbitals. This is in particular useful for probing properties that depend on fine details of the electrons close to the nuclei, such as nuclear magnetic resonance effects [PM01], but comes at the price of a more complex formalism, the response computations in particular being much more cumbersome.

Also notable is the "black-box" approach used in [HGH98; GTH96], where a very simple functional form for the local and nonlocal part of the pseudopotential is postulated, and its parameters adjusted by least-squares fitting to a number of desirable properties, such as the eigenvalues ε_i and norm conservation.

Despite being extremely successful in practice, the various pseudopotential methods remain uncontrolled approximations which are not systematically improvable. For a given density functional, they are often the major source of variation between computational codes [LBBBB+16]. Their error analysis is still underdevelopped; see [CM15] for the existence of norm-conserving pseudopotentials reproducing optimally the first-order Stark effect, and [Dup17] for an analysis of the PAW method for a particular one-dimensional system.

II.7. Solving the discretized problem

Once discretized, the Kohn-Sham problem takes the form

$$\inf_{\langle \phi_i, \phi_j \rangle = \delta_{ij}} E(\phi_1, \dots, \phi_N) \tag{17}$$

where the ϕ_n now belong to a finite-dimensional space $\mathbb{R}^{N_\mathrm{b}}$ (assuming an orthonormal basis for simplicity). The Euler-Lagrange equations for this problem are, up to rotation of the orbitals,

$$H_\Phi \phi_i = \lambda_i \phi_i. \tag{18}$$

This can be seen as either an orthogonality-constrained minimization problem (17), or a nonlinear eigenvector problem (18). Accordingly, there are two main classes of methods to solve this set of equations.

II.7.1. Direct minimization. The first class, *direct minimization* methods, are based on classical methods of continuous optimization: gradient descent, conjugate gradients, quasi-Newton methods, etc. The main obstacle is the constraints $\langle \phi_i, \phi_j \rangle = \delta_{ij}$, which couple together all the orbitals. Fortunately, the matrix set

$$\mathcal{M} = \{\phi = (\phi_1, \dots, \phi_N) \in \mathbb{R}^{N_\mathrm{b} \times N}, \langle \phi_i, \phi_j \rangle = \delta_{ij}\}$$

has the structure of a Riemannian manifold, called the Stiefel manifold. The geometry of this constraint set is particularly nice and enables efficient computations [AMS09].

The simplest iteration is the following projected gradient scheme. For a given $\Phi_n \in \mathcal{M}$, form the gradient $\nabla E(\Phi_n) = H_{\Phi_n} \Phi_n$. Project this gradient onto the tangent space of $\mathcal{M}$ at Φ_n, obtaining

$$P_{T_{\Phi_n}\mathcal{M}}(\nabla E(\Phi_n)) = H_{\Phi_n}\Phi_n - \Phi_n(\Phi_n^* H_{\Phi_n}\Phi_n)$$

where $(\Phi_n^* H_{\Phi_n}\Phi_n)$ is the matrix of the Hamiltonian in the subspace Φ_n. Then take a fixed step $\alpha > 0$ in that direction:

$$\widetilde{\Phi}_{n+1} = \Phi_n - \alpha P_{T_{\Phi_n}\mathcal{M}}(\nabla E(\Phi_n)),$$

and orthogonalize $\widetilde{\Phi}_{n+1}$ with a Löwdin scheme to retract back to $\mathcal{M}$:

$$\Phi_{n+1} = \widetilde{\Phi}_{n+1}\left(\widetilde{\Phi}_{n+1}^* \widetilde{\Phi}_{n+1}\right)^{-1/2}.$$

This scheme is illustrated in Figure 2.

It is easy to see that this scheme is such that $E(\Phi_{n+1}) \leq E(\Phi_n)$ if $\alpha > 0$ is small enough, because $\Phi_{n+1} = \widetilde{\Phi}_{n+1} + O(\alpha^2)$, so that

$$E(\Phi_{n+1}) = E(\Phi_n) - \alpha \| P_{T_{\Phi_n}\mathcal{M}}(\nabla E(\Phi_n)) \|^2 + O(\alpha^2).$$

Accordingly, the convergence of this method can be proven under second-order non-degeneracy conditions [AMS09], or directly thanks to the Łojasiewicz inequality, using the analytic properties of the objective function and constraints [**15**].

Variants of that algorithm can be obtained by the following modifications, the first three of which are classical in all first-order optimization algorithms [NW06]

(1) Preconditioning the algorithm (see Section II.7.4).

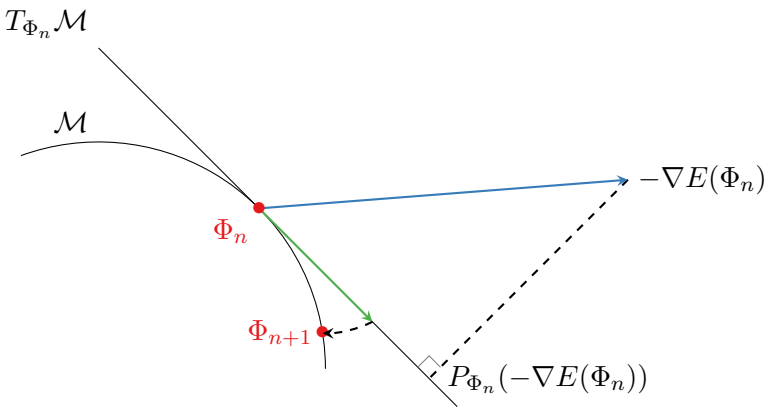


FIGURE 2. Schematic view of the projected gradient algorithm.

(2) Adapting the descent direction by using past iterates, with a scheme such as the conjugate gradient method, or quasi-Newton methods such as the L-BFGS algorithm.
(3) Using an efficient linesearch. For purely gradient methods, the most efficient linesearch seems to be the Barzilai-Borwein method, although it lacks robustness. For methods with better search directions such as L-BFGS, a simple backtracking method often yields the best results.
(4) Varying the orthogonalization method. There is a surprisingly large number of such schemes, among others Gram-Schmidt (classical, modified, or repeated), QR factorization (computed using Householder reflections, Givens rotations or a Cholesky factorization), Löwdin orthogonalization (computed either as above, or through a singular value decomposition) [TBI97]. These methods vary in stability and efficiency (with, roughly, Cholesky-based QR being the most efficient but least stable, and Löwdin orthogonalization being the most stable but least efficient).

II.7.2. Self-consistent field (SCF) algorithm. This class of methods is based on the Euler-Lagrange equations

$$H_\Phi \phi_i = \lambda_i \phi_i.$$

If H did not depend on Φ, this equation would simply be a linear eigenvalue problem, and the minimum of the energy would be achieved by setting the $(\phi_i)_{i=1,\dots,N}$ to be equal to the eigenvectors associated with the first N eigenvalues of H. Let us denote by $\mathcal{A}(H) = (\phi_i)_{i=1,\dots,N}$ this Aufbau mapping, and reformulate the Euler-Lagrange equations as

$$\Phi = \mathcal{A}(H_\Phi).$$

It is natural to try to solve the self-consistent equations by the simple fixed point algorithm

$$\Phi_{n+1} = \mathcal{A}(H_{\Phi_n}).$$

This corresponds to "freezing" the mean field, solving the system of independent particles in the mean-field Hamiltonian, and updating the mean field. Since H only depends on the

density $\rho_\Phi = \sum_{i=1}^N |\phi_i|^2$ (see (10)), this can also be reformulated as

$$\rho_{n+1} = \rho_{\mathcal{A}(H_{\rho_n})}.$$

This is called the self-consistent field (SCF) iteration. While this iteration is numerically found to converge for some small systems, in general it instead overshoots the solution and oscillates. This can be made precise in the case of the HF or rHF model, where the fact that the energy depends quadratically on the density matrix forces the iterates to oscillate between at most two states [CLB00b][**15**]. For the non-quadratic Kohn-Sham model, the behavior is more complicated, supporting cycles of higher periodicity or even chaotic behavior.

The systematic overshooting suggests a damped scheme

$$\rho_{n+1} = \rho_n + \alpha \left(\rho_{\mathcal{A}(H_{\rho_n})} - \rho_n \right),$$

where $\alpha > 0$ is used to interpolate between ρ_n and $\rho_{\mathcal{A}(H_{\rho_n})}$. Let γ_0 be a density matrix associated with ρ_0 (its existence for a wide class of densities is ensured by the N-representability theorem described in Section II.3.1). Then, because the density depends linearly on the density matrix, $\rho_n = \rho_{\gamma_n}$, where

$$\gamma_{n+1} = \gamma_n + \alpha \left(\gamma_{\mathcal{A}(H_{\rho_n})} - \gamma_n \right)$$

and so

$$E(\gamma_{n+1}) = E(\gamma_n) + \alpha \operatorname{Tr}(H_{\rho_n}(\gamma_{\mathcal{A}(H_{\rho_n})} - \gamma_n)) + O(\alpha^2).$$

Since

$$\gamma_{\mathcal{A}(H)} = \operatorname*{arg\,min}_{0 \le \gamma = \gamma^* \le 1, \operatorname{Tr}\gamma = N} \operatorname{Tr}(H\gamma),$$

the energy $E(\gamma_n)$ is decreasing, and so γ_n (and therefore ρ_n) can be proven to converge for α small enough, under relatively general assumptions. In practice, when the explicit computation of γ is possible (which is the case when using for instance gaussian basis functions, but not in the plane wave method), then α can be obtained by a linesearch, resulting in the Optimal Damping Algorithm (ODA) [CLB00a; Can01]. This simple iteration, however, being a stationary iterative method, is slow to converge.

As in the case of gradient descent for direct minimization, this damped (or *mixing*) algorithm is the basic building block of various methods, that precondition (see Section II.7.4) or accelerate it. The most successful acceleration method is known variously as Anderson acceleration [And65], DIIS or Pulay mixing [Pul82]. This method has many avatars, all based on accelerating the convergence of an iteration $x_{n+1} = g(x_n)$ by combining past iterations $g(x_1), \dots, g(x_n)$ to try to minimize some measure of error, the simplest being

$$\begin{aligned} x_{n+1} &= \sum_{m=1}^n \alpha_m g(x_m) \\ \alpha &= \operatorname*{arg\,min}_{\sum_{m=1}^n \alpha_m = 1} \|g(x_m) - x_m\|^2 \end{aligned}$$

This can be seen as a nonlinear version of the GMRES method, which tries to solve a linear system $Ax = b$ by minimizing $\|Ax - b\|$ in the Krylov subspace $\{b, Ab, \dots, A^n b\}$.

When $g(x) = x + (Ax - b)$, the two methods are in fact equivalent [WN11; RS11]. This method is found to spectacularly improve the convergence of the simple damping method, and is the method of choice in many computational codes. However, it is not very robust, unless specific care is taken.

II.7.3. Comparison. At first glance, direct minimization and SCF algorithms look very different. However, we have not yet specified how to solve the eigenvalue problem. While for small systems one can use direct algorithms (such as the QR eigenvalue algorithm), for larger systems this is not feasible and one has to resort to iterative eigensolvers [BDDRV00; Saa11]. Direct minimization methods can be specialized to the case where

$$E(\Phi) = \sum_{i=1}^{N} \langle \phi_i, H\phi_i \rangle$$

for some Hamiltonian H, in which case they turn into eigensolvers for the N lowest eigenvalues of H. In fact, state-of-the-art solvers can be seen as refinement of the humble block gradient method presented above. An example is the Locally-Optimal Block Preconditioned Conjugate Gradient (LOBPCG) algorithm [Kny01]: the Rayleigh-Ritz principle is used to both enforce the orthogonality and select optimal stepsizes using the gradient information ("locally optimal"), preconditioning improves the gradient information ("preconditionned"), and the information from the previous iterate is also included in the search direction ("conjugate"). Such methods have a number of advantages compared to clasical Krylov methods such as the Lanczos method, because they work on blocks of vectors at the same time. This increases the ratio of floating-point operations per memory access, enables easier parallelization, and allow to reuse initial guesses coming from the previous SCF step.

In light of this, direct minimization methods can be interpreted as fusing the two loops in a SCF algorithm based on an iterative eigensolver. Indeed, if the iterative eigensolver is performed using a simple gradient descent and stopped after just one iteration, then the simple SCF algorithm is identical to a direct minimization.

It is hard to reach a general conclusion on the merits of both approaches, as they largely depend on the particular choice of method variant, parameters and implementation. In particular, one deciding factor is the relative cost of the various operations involved. It is here useful to consider extreme cases. First, in quantum chemistry, one often uses small basis sets, but expensive functionals: the full self-consistent Hamiltonian matrix H_Φ is relatively small (and therefore can be diagonalized quickly), but expensive to build. A SCF algorithm is able to extract the maximum amount of information from H_Φ, and is therefore more efficient than direct minimization. At the other extreme, the Gross-Pitaevskii equation has only one orbital, with a simple (quadratic) nonlinearity (see Section V.2). There, the basis set is large and constructing the Hamiltonian explicitly is not possible. Since the Hamiltonian has to be diagonalized iteratively anyway, the SCF method does not appear to offer significant advantages there. KSDFT in a condensed-matter setting seems to sit somewhere between the two, which relatively large basis sets, but many orbitals. Both methodologies, when properly implemented, seem to yield reasonably similar results, although a detailed understanding of this is lacking at present. This is a direction we are currently investigating with Eric Cancès and Gaspard Kemlin.

II.7.4. Preconditioning. The efficiency of iterative methods depends crucially on the conditioning of the problem. In the direct minimization approach, the relevant conditioning is the condition number of the Hessian of the Lagrangian on the tangent space to the minimizer. Preconditioning of the direct minimization method is not very well-developed in the literature (a problem I intend to work on, see Chapter VI). The rest of this section therefore focuses on the SCF method.

The linear eigenproblem. First consider the problem of solving the linear eigenproblem with an iterative method. Although the theory of preconditioning for eigenproblems is less well-studied than that for linear systems, it is by now established [Kny98; Saa11]. We consider for simplicity the problem of computing the non-degenerate smallest eigenvalue of a Hermitian matrix A with eigenvalues $\lambda_1 < \lambda_2 \leq \cdots \leq \lambda_N$, although the analysis extends to block methods that compute several eigenpairs. Consider the simple iteration

$$x_{n+1} = \frac{x_n - \alpha(Ax_n - \langle x_n, Ax_n\rangle x_n)}{\|x_n - \alpha(Ax_n - \langle x_n, Ax_n\rangle x_n)\|},$$

which is a projected gradient descent for the Rayleigh quotient $\langle x, Ax\rangle$ on the sphere $\{x \in \mathbb{C}^N, \|x\| = 1\}$. When α is small enough, this method systematically decreases the Rayleigh quotient, and therefore it can be shown that x_n converges to an eigenvector of A. Generically, the convergence will be to an eigenvector associated with the first eigenvalue of A, and

$$\|Ax_n - \langle x_n, Ax_n\rangle x_n\| \leq C \left(\max_{\lambda \in \sigma(A) \neq \lambda_1} |1 - \alpha(\lambda - \lambda_1)| \right)^n$$

for some $C > 0$. Optimizing this over α, we get $\alpha = \frac{2}{\lambda_1 + \lambda_N}$ and

$$\|Ax_n - \langle x_n, Ax_n\rangle x_n\| \leq C \left(\frac{\kappa - 1}{\kappa + 1} \right)^n$$

The number

$$\kappa = \frac{\lambda_N - \lambda_1}{\lambda_2 - \lambda_1} \geq 1$$

is the condition number of the problem. When κ is large, the number of iterations needed to achieve a given accuracy becomes proportional to κ. Accelerated methods (such as Chebyshev acceleration, conjugate gradients or Krylov methods) will instead have a number of iterations proportional to $\sqrt{\kappa}$ [Saa11].

When A is the discretization of a Schrödinger operator $-\frac{1}{2}\Delta + V$, as the discretization is refined, the largest eigenvalue λ_N grows (for instance, as $O(h^{-2})$ for a finite differences discretization of step h). This deteriorates κ and so the convergence of the algorithm. The source of this behavior is that the residual $r = Ax_n - \langle x_n, Ax_n\rangle x_n$ is not an accurate representation of the error $e = x_n - \lim_{n\to\infty} x_n$. Asymptotically, one has the residual-error relationship

$$r \approx (A - \lambda_1)e,$$

which makes r heavily skewed towards the high-energy eigenmodes of A. The remedy to this issue is well-known: one has to precondition the residual, for instance through

$$x_{n+1} = \frac{x_n - \alpha P^{-1}(Ax_n - \langle x_n, Ax_n\rangle x_n)}{\|x_n - \alpha P^{-1}(Ax_n - \langle x_n, Ax_n\rangle x_n)\|}$$

where P is a preconditionner. Then, the convergence rate of the algorithm depends on the conditionning of

$$(1 - x_1 x_1^*)P^{-1}A(1 - x_1 x_1^*)$$

restricted to the subspace orthogonal to the first eigenvector x_1. We then need to choose P to ensure that this operator remains well-conditioned independently of the discretization parameters. This is the basis for the scheme described in [PTAAJ92], which uses for P an operator that behaves like $(1 - \Delta)^{-1}$. This is easy to achieve in a plane wave basis, where the Laplacian is diagonal.

The SCF algorithm. In the SCF algorithm, the conditioning of the Jacobian matrix of $\rho_{\text{in}} \mapsto \rho_{\mathcal{A}(H_{\rho_{\text{in}}})}$ determines the convergence rate. There are two sources of potential ill-conditioning here: small gaps between occupied and virtual states make the Jacobian of $\mathcal{A}$ diverge, and the Coulomb interaction in the Hartree potential makes the Jacobian of H diverge. The interaction between these two phenomena is not trivial. To analyze it, it is useful to note that the Jacobian of interest is related to the dielectric response of the system. The behavior of this operator depends on the physical nature of the system; it is be analyzed for periodic systems in Section III.3.4, with the conclusion that insulating systems do not need preconditioning, and metallic systems can be preconditioned efficiently with the Kerker scheme.

II.8. Derived properties

In the preceding sections, we have focused on obtaining the electronic energy for a single external potential V_{ext}. This information by itself is not very interesting. However, once this is known, a number of properties can be derived. For instance, the equilibrium configuration of a molecule can be obtained by minimizing the total energy over all possible configurations. The forces on atoms can be obtained as the first derivative of the energy with respect to the atomic positions, enabling the use of molecular dynamics simulations. Similarly, the stress in a periodic crystal can be computed from the first derivative of the total energy per unit volume with respect to the cell parameters.

More advanced properties can be obtained from other types or higher derivatives. For instance, the phonon spectrum describing the vibration modes of atoms in a crystal can be computed from the Fourier transform of the second derivatives of the energy with respect to the atomic positions. Computing the polarizability of a molecule requires computing the derivative of the dipole moment (itself a function of the density) with respect to the external field. Magnetic properties can also be obtained in a similar way.

This task is greatly helped by the Hellmann-Feynman theorem. In its general form it states that, when computing derivatives of the energy, the derivatives of *variational* quantities do not have to be explicitly considered. Consider the function

$$F(y) = \inf_{x \in X} E(x, y)$$

where we assume for simplicitly that E is smooth, and strongly convex with respect to the first variable on a finite-dimensional vector space X for all values of y. Then $x^*(y) = \arg\min_{x\in X} E(x,y)$ is a smooth function of y, satisfying $\frac{\partial E}{\partial x}(x^*(y),y) = 0$. It follows that

$$\frac{dF}{dy} = \frac{\partial E}{\partial x}\cdot\frac{dx^*}{dy} + \frac{\partial E}{\partial y} = \frac{\partial E}{\partial y}.$$

This extends to the case where x is constrained to vary on a manifold $\mathcal{M}$: by the first-order optimality conditions, $\frac{\partial E}{\partial x}$ vanishes on the tangent space $T_x\mathcal{M}$ to which $\frac{dx^*}{dy}$ belongs, and the result

$$\frac{dF}{dy} = \frac{\partial E}{\partial y}$$

also holds. This includes as a special case the derivative of an eigenvalue (minimum of $\langle x, Ax\rangle$ on the unit sphere) with respect to the matrix, which was the original formulation of the Hellmann-Feynman theorem.

In the context of KSDFT, this theorem means that it is very easy to compute the gradient of the electronic energy

$$F(V_{\text{ext}}) = \inf_{\langle\phi_i,\phi_j\rangle=\delta_{ij}} E(\phi_1,\dots,\phi_N;V_{\text{ext}})$$

$$E(\phi_1,\dots,\phi_N,V_{\text{ext}}) = \int_{\mathbb{R}^3}\left(\sum_{i=1}^N \frac{1}{2}|\nabla\phi_i|^2 + V_{\text{ext}}\rho_\phi\right) + \frac{1}{2}\int_{\mathbb{R}^6}\frac{\rho_\phi(x)\rho_\phi(x')}{|x-x'|}dxdx' + E_{\text{xc}}(\rho_\phi)$$

with respect to the external potential:

$$\nabla F = \nabla_{V_{\text{ext}}} E = \rho,$$

where ρ is the minimizing density.

Other types of derivatives are more involved, and require the derivatives of the orbitals ϕ_i with respect to V_{ext}. This is done with density functional perturbation theory, which computes the response of the space $\text{Span}(\phi_i)_{i=1,\dots,N}$ to a change in external potential. We refer to Chapter III.3 for a full derivation in the case of periodic systems at finite temperature, and to [BDGDCG01] (solid-state) and [NRS18] (quantum chemistry) for methods and applications.

CHAPTER III

Periodic quantum systems and defects

In this chapter, I study the modeling of crystals and their defects at the quantum level. I introduce the models used, then outline my contributions to the numerical analysis of the supercell method ([**3**], Section III.2), and to the theoretical analysis of response properties of crystals, in particular the static response of the rHF model in presence of defects ([**2**], Section III.3), and the dynamical response of independent electrons to a uniform electric field ([**1**], Section III.4).

III.1. The rHF model for periodic systems

III.1.1. Thermodynamic limits. We consider the modeling of crystals. A real crystal, even if perfectly pure, necessarily has a finite extension. A physical description of a finite crystal must also model its boundary, which can give rise to boundary effects (for instance surface reconstruction). We are here instead interested in bulk characteristics: for instance, what is the crystal structure of pure silicon? What is its density? To ask this question is to perform implicitly a thermodynamic limit, i.e. to consider a large system and compute its renormalized properties in the limit, ignoring possible surface effects.

In the Born-Oppenheimer approximation, specifying a crystal is specifying a lattice

$$\mathcal{R} = \{n_1 a_1 + n_2 a_2 + n_3 a_3, n_1, n_2, n_3 \in \mathbb{Z}\},$$

where the $(a_i)_{i=1,2,3} \in \mathbb{R}^3$ are the basis vectors, and a $\mathcal{R}$-periodic positive measure μ representing the nuclear charge density. The most natural way to perform the thermodynamic limit is to restrict $\mu_L = \mathbb{1}(\Omega_L)\mu$ to some domain Ω_L of linear size L containing N_L nuclei. We then solve the Schrödinger equation for N_L electrons in $\mathbb{R}^3$ subject to the potential $V_L = -\mu_L * \frac{1}{|x|}$ to obtain the energy E_L. Finally, we compute

$$\lim_{L\to\infty} \frac{E_L}{N_L}$$

as L tend to infinity. For the full Schrödinger equation with Coulomb interaction, it was proved in the landmark paper [Fef85] that the limit above exists. Due to the complexity of that equation, however, this result is quite weak; for instance, it does not ensure that, for fixed periodic nuclei, the resulting electronic density will be periodic. Indeed, symmetry breaking in the form of Wigner crystallization is known to occur in Jellium (where nuclei are modeled by a uniformly charged background) [GV05].

III.1.2. The thermodynamic limit of the rHF model. For the mean-field models mentioned in Chapter II, a systematic program to study the thermodynamic limit was undertaken in the late '90s by Catto, Le Bris and Lions, summarized in [LBL05]. In particular, the convexity of the rHF model forbids symmetry breaking, and it can be proven in this setting that the density ρ_L converges to a periodic density ρ [CLBL01, Theorem 2.2]. We now describe the limit equation satisfied by ρ.

The first difficulty in formulating the limit model is that the nuclear potential $V_L = -\mu_L * \frac{1}{|x|}$ does not make sense in the limit $L \to \infty$, because $\frac{1}{|x|}$ is not integrable at infinity. This is physically obvious, as we are summing the electrostatic potential of an infinite array of positive charges, which must be compensated by the negative charge of

the electrons. This can be remedied by grouping the nuclear potential together with the Hartree potential $\rho_L * \frac{1}{|x|}$. The total charge density $\rho_L - \mu_L$ is then globally neutral, and the total potential $(\rho_L - \mu_L) * \frac{1}{|x|}$ has a limit when $L \to \infty$ and Ω_L is taken to be a cube. The limit energy depends on the particular shape Ω_L chosen for the thermodynamic limit through surface terms. We discard these since we are interested only in bulk properties.

The limit problem, justified mathematically in [CLBL01] (see also [CDL08, Theorem 1]), is most conveniently formulated in terms of the density matrix γ, a $\mathcal{R}$-periodic operator, and the $\mathcal{R}$-periodic density $\rho(x) = \gamma(x, x)$:

$$\left\{ \begin{array}{rl} \gamma & = \mathbb{1}\left(-\frac{1}{2}\Delta + V \leq \varepsilon_F\right) \\ \int_\Gamma \rho & = N_{\text{el}} \\ -\Delta V & = 4\pi(\rho - \mu), \qquad V \ \mathcal{R}\text{-periodic} \end{array} \right. \tag{19}$$

where N_{el} is the number of electrons per unit cell Γ. With appropriate modifications to add an exchange-correlation potential, this is the equation that is solved in practice in Kohn-Sham simulations of materials.

III.1.3. Thermodynamic limit with the supercell method. It is instructive to reproduce the above analysis in a very simplified case. First, we completely neglect the electron-electron interaction: we simply consider N_{el} independent electrons per unit cell in a given potential V. Second, we use the supercell approach: instead of restricting the crystal to a finite-size Ω_L embedded in vacuum, we consider a supercell

$$\Gamma_L = \{x_1 a_1 + x_2 a_2 + x_3 a_3, x_1, x_2, x_3 \in [0, L)^3\},$$

with periodic boundary conditions (and therefore the topology of a torus). This approach is the one most often used in practice for the simulation of materials. This supercell contains L^3 copies of the unit cell of the crystal, and therefore contains $L^3 N_{\text{el}}$ electrons, where N_{el} is the number of electrons per unit cell. These electrons will occupy the first energy levels of the Hamiltonian $H_L = -\Delta + V$ with periodic boundary conditions in Γ_L.

Compared to the method above of embedding the system in vaccum, the great advantage of the supercell method is that it preserves translational symmetry: if $R \in \mathcal{R}$ and τ_R is the operator of translation with vector R, then $H_L \tau_R = \tau_R H_L$. The spectral properties of H_L can therefore be analyzed using the counterpart of Fourier theory for discrete translation symmetry, Bloch-Floquet theory.

The potential V is periodic: for all $R \in \mathcal{R}$, we have $V(\cdot - R) = V(\cdot)$. We introduce the reciprocal lattice

$$\begin{aligned} \mathcal{R}^* &= \{K \in \mathbb{R}^3 \,|\, \forall R \in \mathcal{R}, K \cdot R \in 2\pi\mathbb{Z}\} \\ &= \{n_1 b_1 + n_2 b_2 + n_3 b_3 \,|\, (n_1, n_2, n_3) \in \mathbb{Z}^3\}, \end{aligned}$$

where the vectors b_i are defined by $b_i \cdot a_j = 2\pi\delta_{ij}$. This is the lattice of Fourier coefficients of $\mathcal{R}$-periodic functions. Let

$$L^2_{\text{per}} = \{f \in L^2_{\text{loc}}, f \text{ is } \mathcal{R} - \text{periodic}\}.$$

Any $V \in L^2_{\text{per}}$ can be uniquely decomposed in Fourier series as

$$V(x) = \sum_{K \in \mathcal{R}^*} c_K(V) e^{iKx},$$

the convergence holding in L^2_{per}.

It is instructive to compute the matrix representation of H_L in the basis of plane waves (Fourier modes). Let

$$e_q(x) = \frac{1}{\sqrt{|\Gamma_L|}} e^{iq\cdot x}.$$

This function is compatible with the periodic boundary conditions when q belongs to the supercell reciprocal lattice $\frac{1}{L}\mathcal{R}^*$.

The sublattice $\mathcal{R}^*$ of $\frac{1}{L}\mathcal{R}^*$ induces a natural decomposition: any $q \in \frac{1}{L}\mathcal{R}^*$ can be uniquely decomposed as $k + K$, where $K \in \mathcal{R}^*$ and $k \in \mathcal{B}_L$, where

$$\mathcal{B}_L = \left\{ \frac{n_1}{L} b_1 + \frac{n_2}{L} b_2 + \frac{n_3}{L} b_3, \quad n_1, n_2, n_3 \in \{0, 1, \ldots, L-1\}^3 \right\}.$$

For $q = k + K$ and $q' = k' + K'$, we then have

$$\langle e_q, H_L e_{q'} \rangle = \left(\frac{1}{2}|k+K|^2 \delta_{K,K'} + c_{K-K'}(V) \right) \delta_{k,k'}$$

and we obtain the remarkable phenomenon that H_L is block-diagonal in the $k+K$ representation. There are L^3 blocks, one for each $k \in \mathcal{B}_L$, and the basis functions for each block are $(e_{k+K})_{K\in\mathcal{R}^*}$. This suggests grouping them as the Fourier coefficients of a function, which corresponds to making the Bloch wave ansatz

$$\psi_k(x) = e^{ikx} u_k(x),$$

where u_k is $\mathcal{R}$-periodic. We then have

$$H_L \psi_k = e^{ikx} \left(\frac{1}{2}(-i\nabla + k)^2 + V \right) u_k$$

and the eigenvalues of H_L on the supercell can be obtained by solving the following L^3 eigenvalue problems:

$$\left(\frac{1}{2}(-i\nabla + k)^2 + V \right) u_{nk} = \varepsilon_{nk} u_{nk}$$

for $\mathcal{R}$-periodic functions u_{nk} and $k \in \mathcal{B}_L$. Note that this block-diagonalization reduces an eigenvalue problem on the supercell Γ_L to L^3 eigenvalue problems on the unit cell Γ, a large reduction in complexity.

Since there are $L^3 N_{\mathrm{e}}$ electrons in the system, the occupied states are the $L^3 N_{\mathrm{e}}$ first eigenstates of H_L. Letting ε_F^L be the $(L^3 N_{\mathrm{e}})$-th eigenstate of H_L, assumed non-degenerate, we have

$$\sum_{k\in\mathcal{B}_L, n\in\mathbb{N}} \mathbb{1}(\varepsilon_{nk} \le \varepsilon_F^L) = L^3 N_{\mathrm{el}}, \tag{20}$$

and the total energy is

$$E^L = \sum_{k\in\mathcal{B}_L, n\in\mathbb{N}} \varepsilon_{nk} \mathbb{1}(\varepsilon_{nk} \le \varepsilon_F^L). \tag{21}$$

We now take the thermodynamic limit of these expressions by computing the limit of the energy per unit cell, $\frac{E^L}{L^3}$, as L goes to infinity. Recognizing Riemann sums, we obtain

formally

$$\lim_{L\to\infty} \frac{E^L}{L^3} = \frac{1}{|\mathcal{B}|} \sum_{n\in\mathbb{N}} \int_{k\in\mathcal{B}} \varepsilon_{nk} \mathbb{1}(\varepsilon_{nk} \leq \varepsilon_F) dk,$$

where the *Fermi level* ε_F is obtained implicitly through

$$\frac{1}{|\mathcal{B}|} \sum_{n\in\mathbb{N}} \int_{k\in\mathcal{B}} \mathbb{1}(\varepsilon_{nk} \leq \varepsilon_F) dk = N_{\mathrm{el}}$$

and where the Brillouin zone $\mathcal{B}$ is a unit cell of the reciprocal lattice $\mathcal{R}^*$. This is a specialization of the periodic rHF model (19) in the case without interaction.

III.1.4. The Bloch transform. The set $(\varepsilon_{nk})_{k\in\mathcal{B}}$ for a given n is called a *band*, and the $(\varepsilon_{nk})_{n\in\mathbb{N},k\in\mathcal{B}}$ collectively form the *band structure*. The band structure characterizes the spectrum of $H = -\frac{1}{2}\Delta + V$ on the whole space $\mathbb{R}^3$: for instance,

$$\sigma(H) = \bigcup_{k\in\mathcal{B},n\in\mathbb{N}} \{\varepsilon_{nk}\}. \tag{22}$$

More formally, the above statements can be written using the Bloch transform $\mathcal{U}$, which is to Bloch waves what the Fourier transform is to plane waves. This transform is defined for Schwartz functions by [RS72, Vol. 4]:

$$(\mathcal{U}w)_k(x) = \sum_{R\in\mathcal{R}} w(x+R)\, e^{-ik\cdot(x+R)},$$

with inverse

$$(\mathcal{U}^{-1}u)(x) = \frac{1}{|\mathcal{B}|} \int_{\mathcal{B}} e^{ik\cdot x} u_k(x)\, dk.$$

The Bloch transform expresses any function w as a linear combination of Bloch waves $e^{ikx}u_k(x)$, with $u_k(x)$ periodic in x. Up to a constant factor, the Bloch transform extends to a unitary map from $L^2(\mathbb{R}^3)$ to $L^2(\mathcal{B}, L^2_{\mathrm{per}})$.

We say that a bounded operator $A : L^2(\mathbb{R}^3) \to L^2(\mathbb{R}^3)$ is periodic if, for all $R \in \mathcal{R}$,

$$A(w(\cdot - R)) = (Aw)(\cdot - R).$$

If A is periodic, we have seen above that A does not couple Bloch waves with different pseudo-momentum k. More precisely, A is decomposed by the Bloch transform:

$$(\mathcal{U}(Aw))_k(x) = (A_k(\mathcal{U}w)_k)(x),$$

where the operator $A_k = e^{-ikx}Ae^{ikx}$ is a bounded operator on L^2_{per}, called the fiber of A at k. The formalism extends to unbounded operators through the resolvent, and implies for instance that

$$\left(-\frac{1}{2}\Delta + V\right)_k = \frac{1}{2}(-i\nabla + k)^2 + V$$

when V is a periodic potential. The fact that periodic operators do not couple different k then leads to the decomposition (22) of the spectrum.

Using the Bloch transform, the previous thermodynamic limit can now be reformulated as

$$\lim_{L\to\infty} \frac{E^L}{L^3} = \underline{\mathrm{Tr}}(H\gamma),$$

with

$$\gamma = \mathbb{1}(H \leq \varepsilon_F),$$
$$\underline{\mathrm{Tr}}(\gamma) = N_{\mathrm{el}},$$

where, for a periodic, locally trace-class operator A, the trace per unit cell $\underline{\mathrm{Tr}}(A)$ is given by

$$\underline{\mathrm{Tr}}(A) = \lim_{L\to\infty} \frac{1}{L^3} \mathrm{Tr}(\chi_{\Gamma_L} A \chi_{\Gamma_L}) = \frac{1}{|\mathcal{B}|} \int_{\mathcal{B}} \mathrm{Tr}(A_k) dk$$

where $\Gamma_L = L\Gamma$ with Γ a unit cell of the crystal, and χ_Ω is the indicator function of Ω.

III.2. Numerical analysis

III.2.1. Supercell method and numerical integration. In practice, the properties of a crystal are computed by the supercell approach above: we sample the Brillouin zone $\mathcal{B}$ at L^3 points $\mathcal{B}_L$, solving L^3 eigenvalue problems. Since each of these eigenvalue problems is expensive to solve, it is important to keep L minimal, and therefore to ask the question: for a given number of eigenvalue problem solves, what is the optimal sequence of computation to get an approximation of the thermodynamic limit?

We seek to estimate the error in the supercell method of quantities such as

$$\Delta\varepsilon_F(L) = \left|\varepsilon_F^L - \varepsilon_F\right|,$$
$$\Delta E(L) = \left|\frac{E^L}{L^3} - \lim_{L\to\infty} \frac{E^L}{L^3}\right|,$$

with respect to L, where ε_F^L and E^L were defined in (20) and (21). Other measures of error are possible; we focus on these for simplicity. The behavior of the error with respect to L turns out to depend crucially on whether the material under consideration is gapped or not. A periodic crystal is an insulator if there is a gap between the occupied and the unoccupied states, i.e. if

$$\inf_{k\in\mathcal{B}} \varepsilon_{N_{\mathrm{el}}+1,k} > \sup_{k\in\mathcal{B}} \varepsilon_{N_{\mathrm{el}},k}.$$

In the case where there is no gap, the system is metallic.

For an insulator, the formulas above simplify to

$$E^L = \sum_{k\in\mathcal{B}_L} \sum_{n=1}^{N} \varepsilon_{nk}$$

It is then clear that the supercell method is approximating

$$\lim_{L\to\infty} \frac{E^L}{L^3} = \frac{1}{|\mathcal{B}|} \int_{k\in\mathcal{B}} \sum_{n=1}^{N} \varepsilon_{nk} dk$$

by a simple Riemann sum. Naively, we would therefore expect an error of $O(1/L^2)$. However, the integrand is periodic, and Riemann sums are much more efficient than this estimate suggests (see Figure 1) [TW14]. The reason for this is that

$$\frac{1}{L}\sum_{n=0}^{L-1} e^{im\frac{2\pi n}{L}} - \frac{1}{2\pi}\int_0^{2\pi} e^{imx}dx = \begin{cases} 1 & \text{if } m \text{ is a non-zero multiple of } L, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, the Riemann sum method is *exact* on low-degree trigonometric functions. Functions with frequencies multiple of L are aliased on the grid to constant functions, and therefore introduce an error. It follows that the faster the decay of the Fourier coefficients, the faster the convergence of the method [TW14]. In our case, the map $k \mapsto \sum_{n=1}^{N} \varepsilon_{nk}$ is analytic, and so, by a Paley-Wiener argument, one can show that

$$\Delta E(L) \leq Ce^{-\alpha L}$$

for some $C, \alpha > 0$. This was extended in [GL16] to the case of the rHF model.

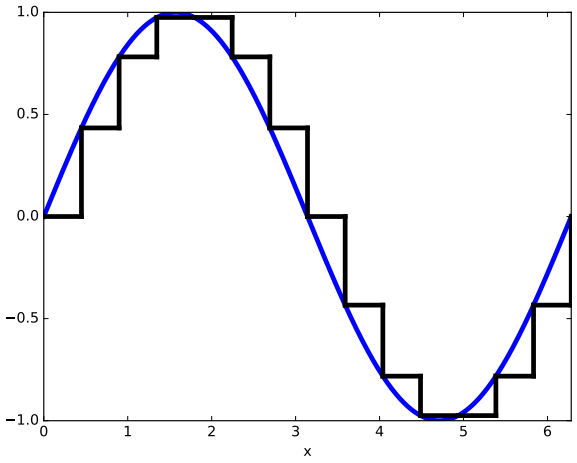


FIGURE 1. Integration of a periodic function by the Riemann sum method. By periodicity, the underestimation of the area under the curve almost cancels with the overestimation, yielding very fast convergence.

III.2.2. The supercell method for metals. This exponential convergence for insulators is borne out by numerical practice: very good results are obtained already with coarse grids such as $L = 4$ or $L = 6$. However, for metals, the convergence is much more erratic, and considerably larger grids are used. The difficulty for metals is that the integrands $k \mapsto \sum_{n\in\mathbb{N}} \mathbb{1}(\varepsilon_{nk} \leq \varepsilon_F)$ and $k \mapsto \sum_{n\in\mathbb{N}} \varepsilon_{nk}\mathbb{1}(\varepsilon_{nk} \leq \varepsilon_F)$ are now discontinuous functions in the Brillouin zone $\mathcal{B}$. Therefore, approximating the integral by a Riemann sum is very inaccurate. The locus of discontinuities is the Fermi surface

$$\mathcal{S} = \cup_{n\in\mathbb{N}}\mathcal{S}_n,$$

where the n-th sheet is

$$\mathcal{S}_n = \{k \in \mathcal{B}, \exists n \in \mathbb{N}, \varepsilon_{nk} = \varepsilon_F\}.$$

The Brillouin zone might have multiple sheets and complex shapes; see for instance http://www.phys.ufl.edu/fermisurface for visualizations of the Fermi surfaces of pure solids.

The convergence of ΔE^L depends on the properties of the band structure around the Fermi surface. In [3], we use the following assumptions:

ASSUMPTION III.2.1. *There are no eigenvalue crossings at the Fermi level:* $\forall n \neq m, \mathcal{S}_n \cap \mathcal{S}_m = \emptyset$;

ASSUMPTION III.2.2. *There are no flat bands at the Fermi level:* $\forall n \in \mathbb{N}, \forall k \in \mathcal{S}_n, \nabla \varepsilon_{nk} \neq 0$.

The violation of these two assumptions can create *van Hove singularities*, which are points of non-smoothness of the density of states. These assumptions are true for most materials, with the notable exception of graphene (which violates the first condition).

Under these assumptions, the Fermi surface is a smooth surface, and we can prove ([3], Theorem 4.5) that

$$\Delta E(L) \leq \frac{C}{L}$$

for some constant $C > 0$, as expected from a Riemann sum of a piecewise smooth function. This convergence is very slow in practice, and a large literature has focused on obtaining schemes with better convergence properties.

The first idea is to use higher-order interpolation: using the values of ε_{nk} sampled at the points of $\mathcal{B}_L$, construct piecewise polynomial interpolants P_{nk} and Q_{nk} which approximate ε_{nk} to order $p+1$ and $q+1$ respectively as $L \to \infty$. Then, compute $\varepsilon_F^{L,q}$ and $E^{L,p,q}$ through

$$\frac{1}{|\mathcal{B}|} \int_{\mathcal{B}} \sum_{n \in \mathbb{N}} \mathbb{1}(Q_{nk} \leq \varepsilon_F^{L,q}) dk = N_{\text{el}}, \tag{23}$$

$$E^{L,p,q} = \frac{1}{|\mathcal{B}|} \int_{\mathcal{B}} \sum_{n \in \mathbb{N}} P_{nk} \mathbb{1}(Q_{nk} \leq \varepsilon_F^{L,q}) dk. \tag{24}$$

For these methods, we proved the following result.

THEOREM III.2.3. *Assume that Assumption III.2.1 is satisfied, and that the interpolants* P_{nk}, Q_{nk} *are local and accurate to order* $p+1$ *and* $q+1$ *respectively. Then there exists* $C > 0$ *such that, for* L *large enough,*

$$|\varepsilon_F^{L,q} - \varepsilon_F| \leq \frac{C}{L^{q+1}},$$

$$|E^{L,p,q} - E| \leq C \left(\frac{1}{L^{p+1}} + \frac{1}{L^{2q+2}} \right).$$

We refer to [3] Theorem 4.5 for the precise assumptions on the interpolants. The surprise here is the appearance of $2q+2$ instead of $q+1$: although the computation of the energy depends on the Fermi level, it can be significantly more accurate. The solution to this paradox is that the error made on the ground state energy is, to leading order, equal to ε_F times the error made on the number of electrons, which is zero from (23).

To implement these methods, one needs to be able to compute analytically quantities of the form

$$\int_{\mathcal{B}} P(k) \mathbb{1}(Q(k) \leq \varepsilon) dk$$

for piecewise polynomials P and Q. This is achievable when $p = q = 1$, but becomes very complex for larger degrees. The most efficient method used in practice is the improved tetrahedron method [BJA94], which is a correction to the method with $p = q = 1$ to achieve results similar to the $p = 2, q = 1$ method.

III.2.3. Smearing. In practice, the most used method for the computation of the properties of metals is to add an artificial *smearing*. This is inspired by systems at finite temperature: a system of N_{el} non-interacting fermions in the canonical ensemble at temperature T is characterized by a density matrix γ obeying the Fermi-Dirac statistics

$$\begin{aligned} \gamma &= f\left(\frac{H - \varepsilon_F}{k_B T}\right) \\ \underline{\mathrm{Tr}}\gamma &= N_{\text{el}} \end{aligned} \tag{25}$$

with

$$f(x) = \frac{1}{1 + e^x}$$

the Fermi-Dirac occupation function[1].

Binding energies in solids are on the order of the electron-volt, equivalent to a $k_B T$ factor of about $10,000$K. Therefore, although temperature effects on electronic properties are relevant to systems under extreme conditions (in the core of giant planets, or in inertial confinement fusion [SSB18]), they are not relevant at room temperature. However, they are convenient numerically: the supercell energy E^L becomes

$$\frac{E^{L,T}}{L^3} = \frac{1}{L^3} \sum_{k \in \mathcal{B}_L, n \in \mathbb{N}} \varepsilon_{nk} f\left(\frac{\varepsilon_{nk} - \varepsilon_F^{L,T}}{k_B T}\right),$$

where the approximate Fermi level $\varepsilon_F^{L,T}$ is chosen to satisfy

$$\frac{1}{L^3} \sum_{k \in \mathcal{B}_L, n \in \mathbb{N}} f\left(\frac{\varepsilon_{nk} - \varepsilon_F^{L,T}}{k_B T}\right) = N_{\text{el}}.$$

When $T > 0$, this is a Riemann sum method on a smooth integrand, and converges exponentially fast with respect to L. This is a standard numerical trick in computations on metals: adding an artificial small temperature helps convergence. The artificial temperature is chosen as large as possible without interfering with the physical results; the widely used ABINIT code has a default value of T of about $3,000$K.

It is also convenient to use other (non-physical) smearing functions than the Fermi-Dirac distribution. Figure 2 shows some of the most commonly used. In particular, the popular Methfessel-Paxton scheme [MP89] replaces the Fermi-Dirac occupation function f by one for which f' has more zero moments, and is therefore a higher-order approximation

[1]Note that there are subtleties involved in performing thermodynamic limits *à la* Catto-Le Bris-Lions at finite temperature: operators representing non-confined particles in the whole space possess continuous spectrum, and density matrices of the form (25) cannot be trace-class if H has continuous spectrum. Physically, this is because a system cannot be kept at finite temperature equilibrium if it is in contact with the vacuum. This issue is not present when considering confined particles, either through Dirichlet boundary conditions, or the supercell method.

of the Dirac distribution. This results in a better approximation of quantities of interest, as the following theorem shows.

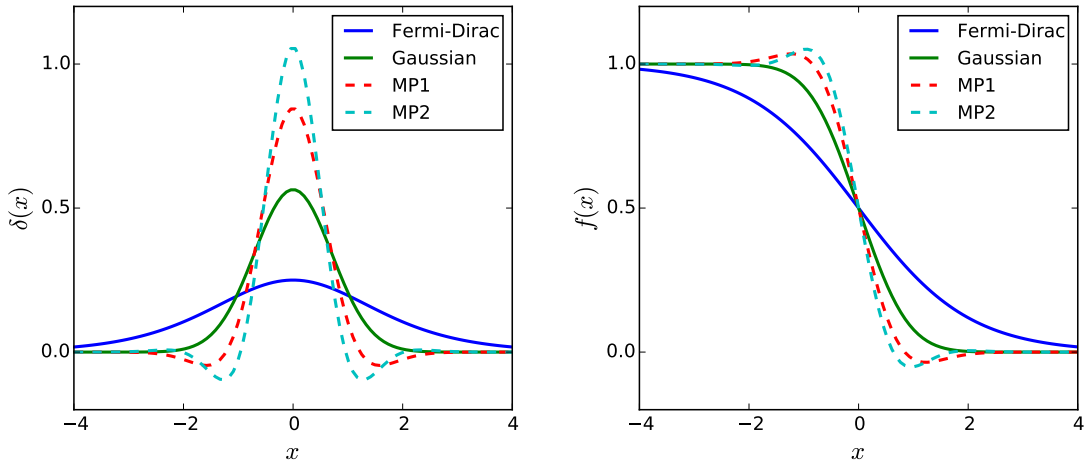


FIGURE 2. Some smearing functions used in practice. Approximation to the Dirac function δ^1 (left), and occupation numbers f^1 (right).

THEOREM III.2.4. *Assume that assumptions III.2.1 and III.2.2 hold, and that f obeys some technical conditions satisfied by all methods in practice (see [3] for details). Assume that f is of order at least p, i.e. that $\int_{\mathbb{R}} x^n f'(x)dx = 0$ for all $1 \leq n \leq p$. Then there are $C > 0, \eta > 0$ such that*

$$\Delta E(L,T) + \Delta\varepsilon_F(L,T) \leq C\Big(T^{p+1} + T^{-4}e^{-\eta TL}\Big).$$

Introducing temperature therefore makes the problem easier to solve (since the integrand is smooth), at the price of an error $O(T^{p+1})$. In principle, by optimizing over T at finite L one can select the appropriate rate of decay to minimize the error for a given computational budget L; in particular, by selecting $L = \frac{1}{T^{1+\varepsilon}}$, one gets a total error proportional to $L^{-\frac{p+1}{1+\varepsilon}}$.

There are two parts in the estimates of this theorem: first, the error made by introducing the artificial temperature, and second the quadrature error. The simplest way to understand the first error is to notice that, if

$$\mathcal{E}(\varepsilon) = \frac{1}{|\mathcal{B}} \int_{\mathcal{B}} \sum_{n\in\mathbb{N}} \varepsilon_{nk} \mathbb{1}(\varepsilon_{nk} \leq \varepsilon) dk,$$

then

$$\mathcal{E}^T(\varepsilon) := \frac{1}{|\mathcal{B}} \int_{\mathcal{B}} \sum_{n\in\mathbb{N}} \varepsilon_{nk} f\left(\frac{\varepsilon_{nk} - \varepsilon}{k_B T}\right) dk = (\mathcal{E} * \delta^T)(\varepsilon),$$

where

$$\delta^T(\varepsilon) = -\frac{1}{k_B T} f'\left(\frac{\varepsilon}{k_B T}\right).$$

Since $\int_{-\infty}^{+\infty} f' = 1$, when $T \to 0$, δ^T tends to the Dirac distribution. More precisely, for all ϕ in the Schwartz space $\mathcal{S}$, we have the expansion

$$-\int_{-\infty}^{+\infty} \delta^T(\varepsilon)\phi(\varepsilon)d\varepsilon = \phi(0) + \cdots + \frac{T^n}{n!}\left(\int_{\mathbb{R}} x^n f'(x)dx\right)\phi^{(n)}(0) + O(T^{n+1}).$$

and so, provided that $\mathcal{E}$ is smooth at ε,

$$\mathcal{E}^T(\varepsilon) = \mathcal{E}(\varepsilon) + O(T^{p+1}).$$

Second, for a given T, the convergence is exponential with respect to L, because the integrand

$$k \mapsto \sum_{n\in\mathbb{N}} \varepsilon_{nk} f\left(\frac{\varepsilon_{nk} - \varepsilon}{k_B T}\right) = \mathrm{Tr}\left(H_k f\left(\frac{H_k - \varepsilon}{k_B T}\right)\right)$$

is analytic in a complex strip around the real axis. However, the width of this complex strip depends on the temperature, which is the origin of the temperature dependence of the exponent $e^{-\eta TL}$. Finally, when the smearing function f is entire (as is the case with the Methfessel-Paxton scheme), then the convergence is super-exponential (see **[3]**).

III.2.4. Perspectives. For insulators, it is clear that the supercell method is the correct choice, and leads to a very fast convergence in practice. For metallic systems, however, it is extremely slow, and more sophisticated methods are needed to keep the computation cost acceptable. We described elementary approaches that only work on the eigenvalues ε_{nk}. By using higher-order interpolants or by artificially smearing the system, one is able to achieve arbitrary high convergence orders. In practice, though, higher-order methods are found not to be competitive, and only methods of order 1 or 2 are used.

The main problem in the interpolation of bands is that they undergo crossings or avoided crossings, at which points they are not smooth. A way to bypass that issue is to use more information than simply the ε_{nk}. For instance, several methods exploit the fact that the local behavior of the ε_{nk} with respect to k can be computed efficiently from perturbation theory (the "$k \cdot p$" method) [PP99]. Others use a reduced basis methodology to solve smaller-size eigenvalue problems [Shi96]. Yet another option is Wannier interpolation, described in Chapter IV. Of these, Wannier interpolation seems to be the most accurate, but the systematic construction of good Wannier functions is not trivial.

Finally, note that these more sophisticated interpolation techniques are most often used in a non-self-consistent way, i.e. for a fixed, converged, Hamiltonian. The coupling to the SCF cycle on the one hand, and to properties (derivatives of the energy) on the other, is not well studied. I intend to pursue this in future work.

III.3. Defects, screening and charge sloshing

We have until now been concerned with a perfectly periodic crystal. While the structure of perfect crystals successfully explains a number of the properties of real crystals (for instance, density, absorption spectrum, Young's modulus), others, such as conductivities, tensile strength or ductility, depend crucially on the deviations of the crystal from its "perfect" equivalent. For instance, steel (which contains by mass around 99% iron and 1% carbon) has very different properties than those of iron: "crystals are like people: it is the defects in them which tend to make them interesting" [Hum79]. Mathematically, the modeling of various types of defects is the next logical step in the complexity ladder

towards realistic systems once periodic systems are understood, and consequently has attracted a large amount of interest, at both the quantum and classical level (see [CLB13] for a review).

In this section, we study the electrostatic response of crystals to defects in the reduced Hartree-Fock approximation. This is interesting in itself, and also to understand the convergence properties of the self-consistent iterations for extended systems. We begin with a phenomenological description in Sections III.3.1 and III.3.1, then state the mathematical results obtained in **[2]** in Sections III.3.3 and III.3.4.

III.3.1. Model and response functions. Imagine putting a free charge Q in the crystal, with a charge distribution $\mu_{\text{def}} = Q\delta_0$. In vacuum, this would create a potential $V(x) = \frac{Q}{|x|}$. In a material, neglecting the movement of the nuclei, electrons will reorganize to this defect, and the total potential V will include the Coulomb interaction created by μ_{def} and the electron response ρ_{def}. There is a large empirical difference in behavior depending on whether charge carriers in the material are mobile or not. In an insulator, electrons are tightly bound to their nuclei, and will only shift slightly away from their equilibrium positions. A simple empirical model is

$$V(x) \approx \frac{Q}{|x|\varepsilon_r}$$

for a material-dependent dielectric constant $\varepsilon_r > 1$. This is the reason why dielectric materials are often inserted in capacitors: by effectively reducing the electric field, they achieve a larger capacitance. By contrast, in metals, electrons are mobile, and flock even from far away towards a positive charge (or move away from a negative charge). At equilibrium, the defect is surrounded by an excess (or depletion) of electrons that nullifies its effect at long range, an empirical model being the Yukawa or screened Coulomb potential

$$V(x) \approx \frac{Q}{|x|} e^{-k|x|}$$

where $1/k$ is the screening length.

How do these effects arise from microscopic theory? We will provide a partial answer in the rHF model of defects. We described in Section III.1 the rHF model for a periodic crystal with a given nuclei distribution μ. Let $\rho_{\text{per}}(x) = \mathbb{1}(-\Delta + V_{\text{per}} \leq \varepsilon_F)(x,x)$ be the electronic density, with V_{per} the unique periodic solution of $-\Delta V_{\text{per}} = 4\pi(\rho_{\text{per}} - \mu)$. Let us now consider a defect, modeled by a local charge density μ_{def}. The model is formulated in terms of the electronic excess density ρ_{def}:

$$\left\{ \begin{array}{rl} \rho_{\text{def}}(x) & = \mathbb{1}(-\Delta + V_{\text{per}} + V \leq \varepsilon_F)(x,x) - \rho_{\text{per}}(x), \\ -\Delta V & = 4\pi(\rho_{\text{def}} - \mu_{\text{def}}), \end{array} \right. \tag{26}$$

where V is the (non-periodic) total potential resulting from the defect. This model was studied in [CDL08], and derived from a supercell thermodynamical limit. It is the one computed in practice for the study of defects in crystalline solids. With these definitions, the picture expected empirically is given in Figure 3.

When μ_{def} is small, or when it is very extended (so that $\mu_{\text{def}} * \frac{1}{|x|}$ is small), one can use the tools of linear response. The linear response of the rHF model to defects was performed first in the electron gas (V_{per} constant) by Lindhard [Lin54], providing the

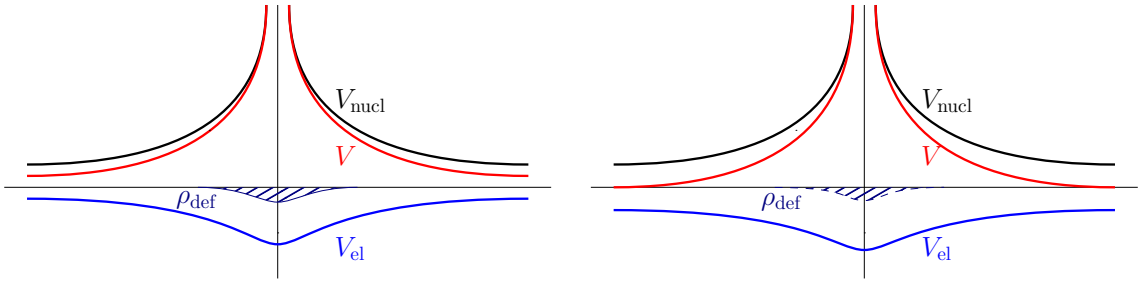


FIGURE 3. Quantities involved in (26), with $-\Delta V_{\text{el}} = 4\pi\rho_{\text{def}}$, $-\Delta V_{\text{nucl}} = -4\pi\mu_{\text{def}}$ and $V = V_{\text{nucl}} + V_{\text{el}}$. The positive nuclear defect potential V_{nucl} locally depletes the electrons, creating a counter-potential V_{el} which reduces the total potential V. In insulators (left panel), the defect charge satisfies $\int_{\mathbb{R}^3} \rho_{\text{def}} < -\int_{\mathbb{R}^3} \mu_{\text{def}}$, and so the total field is long-ranged (partial screening). In metals (right panel), $\int_{\mathbb{R}^3} \rho_{\text{def}} = -\int_{\mathbb{R}^3} \mu_{\text{def}}$ and so it is short-ranged (full screening). Lattice-scale oscillations are not pictured.

first quantum-mechanical understanding of screening in metals. In the '60s, Adler and Weiser extended this to arbitrary periodic potentials [Adl62; Wis63], building upon earlier work of Ehrenreich and Cohen [EC59] that ignored the "local field effects" (the coupling of reciprocal wavectors in the response functions). This was studied mathematically in [CL10].

Formally, linear response proceeds by expanding V to first order in μ_{def}. Denote by χ_0 the independent-particle polarizability operator, defined as the expansion of ρ_{def} to first order in V:

$$\mathbb{1}(-\Delta + V_{\text{per}} + V \leq \varepsilon_F)(x,x) \approx \rho_{\text{per}}(x) + (\chi_0 V)(x).$$

The equation (26) becomes

$$(-\Delta - 4\pi\chi_0)V = -4\pi\mu_{\text{def}},$$

and so

$$V = -4\pi A\mu_{\text{def}}, \tag{27}$$

$$A = (-\Delta - 4\pi\chi_0)^{-1} \tag{28}$$

The properties of the transfer operator A depend crucially on those of χ_0. This latter operator is periodic: it commutes with lattice translations. Therefore, it can be block-diagonalized into fibers $\chi_{0,q}$ in the same way as the Hamiltonian. The operator A is periodic, with fibers $A_q = ((-i\nabla + q)^2 - 4\pi\chi_{0,q})^{-1}$. This is a complicated expression: not only do we have to compute $\chi_{0,q}$, but we also have to invert the operator $((-i\nabla + q)^2 - 4\pi\chi_{0,q})$. A special case is when the external potential is considered to be constant. In this case, the operator A commutes with all translations, and is given by a simple multiplication in Fourier space by the number $(|q|^2 - 4\pi\chi_{0,q})^{-1}$, yielding the Lindhard response function [GV05].

In the general case, if we denote by u_{nk} and ε_{nk} the orthonormal eigenfunctions and eigenvalues of

$$((-i\nabla + k)^2 + V)u_{nk} = \varepsilon_{nk}u_{nk},$$

then by first-order perturbation theory we can compute

$$\chi_{0,q} = \frac{1}{|\mathcal{B}|}\int_{\mathcal{B}} \sum_{n,m\in\mathbb{N}} \frac{\mathbb{1}(\varepsilon_{n,k+q} \le \varepsilon_F) - \mathbb{1}(\varepsilon_{mk} \le \varepsilon_F)}{\varepsilon_{n,k+q} - \varepsilon_{mk}} |\overline{u_{mk}}u_{n,k+q}\rangle\langle\overline{u_{mk}}u_{n,k+q}|dk, \tag{29}$$

the Adler-Wiser formula [Adl62; Wis63].

III.3.2. Insulators and metals. For general periodic systems, the behavior of $\chi_{0,q}$ for large wavelength is very different in insulators and in metals. This can most easily be seen by expressing the operator $\chi_{0,q}$ in the basis of Fourier modes $e_K(x) = \frac{1}{\sqrt{|\Gamma|}}e^{iKx}, K \in \mathcal{R}^*$. We have

$$\langle e_K, \chi_{0,q}e_{K'}\rangle = \frac{1}{|\mathcal{B}|}\int_{\mathcal{B}} \sum_{n,m\in\mathbb{N}} \frac{\mathbb{1}(\varepsilon_{n,k+q} \le \varepsilon_F) - \mathbb{1}(\varepsilon_{mk} \le \varepsilon_F)}{\varepsilon_{n,k+q} - \varepsilon_{mk}} \langle e_K|\overline{u_{mk}}u_{n,k+q}\rangle\langle\overline{u_{mk}}u_{n,k+q}|e_{K'}\rangle dk.$$

For insulators, when q is small, the only possible excitations are between occupied and empty states, i.e. $n \le N_{\text{el}}, m > N_{\text{el}}$ or $n > N_{\text{el}}, m \le N_{\text{el}}$. In this case, when $K = 0$, we have the expansion

$$\langle e_0|\overline{u_{mk}}u_{n,k+q}\rangle = \frac{1}{\sqrt{|\Gamma|}}\langle u_{mk}|u_{n,k+q}\rangle = q \cdot \frac{1}{\sqrt{|\Gamma|}}\langle u_{mk}|\nabla_k u_{n,k}\rangle + O(|q|^2).$$

It follows that we have

$$(\chi_{0,q})_{00} = O(|q|^2), \quad (\chi_{0,q})_{K0} = O(|q|), \quad (\chi_{0,q})_{0K} = O(|q|)$$

and so, separating the "head" $(0,0)$, "wings" $(K,0),(0,K)$ with $K \neq 0$ and "tail" (K,K') with $K, K' \neq 0$, we get the block decomposition

$$(-i\nabla + q)^2 - 4\pi\chi_{0,q} = \begin{pmatrix} O(|q|^2) & O(|q|) \\ O(|q|) & |K+q|^2\delta_{KK'} - 4\pi(\chi_{0,q})_{KK'}. \end{pmatrix}$$

Furthermore, it is easily seen that $\chi_{0,q}$ is nonpositive, so that the "tail" is invertible. By a Schur complement formula, it follows that

$$(A_q)_{00} = ((-i\nabla + q)^2 - 4\pi\chi_{0,q})^{-1}_{00} = \frac{1}{q^T\varepsilon_r q} + O\left(\frac{1}{|q|}\right) \tag{30}$$

for some 3×3 symmetric matrix $\varepsilon_r > 1$.

For metals, by contrast, excitations $u_{mk} \leftrightarrow u_{n,k+q}$ at arbitrary small q are allowed. In fact, when $q \to 0$, the diagonal terms $m = n$ in $(\chi_{0,q})_{00}$ contribute near the Fermi surface; a more careful analysis shows that this term is proportional to the density of states at the Fermi level, which is non-zero for a "regular" metal (one satisfying assumptions III.2.1 and III.2.2). As a result, $(A_q)_{00}$ does not diverge but tends to a finite limit as $q \to 0$.

The result then is that the Fourier transform of the potential response to a local charge density will have a $\frac{1}{|q|^2}$ divergence in insulators, but not in metals. This corresponds to the presence of a long-range Coulomb force in insulators (partially screened by a factor ε_r), but not in metals (fully screened response).

Note that we have here focused only on the behavior at $q = 0$. This is indeed the source of partial screening in insulators, and total screening in finite-temperature systems. For zero-temperature metals, another phenomenon appears: although the response functions do not diverge at $q = 0$, similar to the case of finite-temperature systems, it does not mean that the total potential is exponentially localized, because the discontinuities in k-space produce oscillatory tails. This can be understood most simply by considering the simple case $V_{\text{per}} = 0$, in which case χ_0 is a translation invariant operator characterized by its reciprocal space expression $\chi_0(q)$. This can be analytically computed (see [GV05]), and has a singularity at $q = 2k_F$, where k_F is the Fermi wave vector. This results in long-range oscillations known as Friedel oscillations. In arbitrary periodic systems, the long-range behavior of the total potential depends on the shape of the Fermi surface [RZK66].

It is instructive to consider the above analysis in the framework of the simpler local "Poisson+F" model

$$-\Delta V = 4\pi(\rho^F_{\text{def}}[V] - \mu_{\text{def}})$$

where

$$\rho^F_{\text{def}}[V](x) = F(V(x))$$

with $F : \mathbb{R} \to \mathbb{R}$ a given function. Models of this form include the Thomas-Fermi equation for the electron gas (see Section II.3.1) and the Poisson-Boltzmann equation for ionic liquids. Assuming that μ_{def} is small, we can linearize this equation to

$$-\Delta V - 4\pi\chi_0 V = -4\pi\mu_{\text{def}},$$

where $\chi_0 = F'(0)$. The solution of this equation for $\chi_0 \leq 0$ is

$$V(x) = \frac{Q}{|x|} e^{\sqrt{-4\pi\chi_0}|x|}.$$

This makes the physical interpretation clearer: χ_0 is the susceptibility, and quantifies the amount of charge carriers that are created in the system by a raising of the potential V. When charge carriers are available ($\chi_0 < 0$), total screening results.

The rHF model is more complex to interpret due to the non-locality of the mapping χ_0. The number $(\chi_{0,q})_{KK'}$ represents the response at wavelength $q+K$ of the density a crystal of independent electrons to a perturbation of the potential at wavelength $q+K'$. Lowering the potential at a low wavelength (q small, $K = 0'$) corresponds to raising the Fermi level locally. For insulators, this does not change the occupations: $(\chi_{0,0})_{00} = 0$ (there are no available carriers). For metals, this increases the density locally: $(\chi_{0,0})_{00} < 0$. As in the simple "Poisson+F" model, this results in total screening in the case of metals, but not in the case of insulators. Additional effects compared to the "Poisson+F" model are the appearance of partial screening and Friedel oscillations.

III.3.3. Rigorous results. To make the analysis above rigorous, we need to define the model of defect, justify the linear response (formulas (27) and (29) for the operators A and χ_0), and finally compute their properties precisely. This was done for insulators in [CL10]. There, a homogenization limit was considered: the authors considered a dilute defect charge

$$\mu_{\text{def},\eta}(x) = \eta^3 \mu_{\text{def}}(\eta x)$$

for a localized profile $\mu_{\rm def}$ and η small. They then proved that the rescaled potential $\eta^{-1}V_\eta(\eta^{-1}x)$ converges weakly to the solution of the Poisson equation

$$\operatorname{div}(\varepsilon_r \nabla V) = 4\pi\mu_{\rm def}$$

(compare with (30)).

In [2], I considered the case of metals. To avoid complications related to Friedel oscillations, I used the finite-temperature model

$$\left\{\begin{array}{rl} \rho_{\rm per} & = f(-\Delta + V_{\rm per} - \varepsilon_F)(x,x), \\ \int_\Gamma \rho_{\rm per} & = N_{\rm el}, \\ -\Delta V_{\rm per} & = 4\pi(\rho_{\rm per} - \mu_{\rm per}), \quad V\ \mathcal{R}\text{-periodic}, \end{array}\right. \tag{31}$$

for the periodic system, and

$$\left\{\begin{array}{rl} \rho_{\rm def}(x) & = f(-\Delta + V_{\rm per} + V - \varepsilon_F)(x,x) - \rho_{\rm per}(x), \\ -\Delta V & = 4\pi(\rho_{\rm def} - \mu_{\rm def}) \end{array}\right. \tag{32}$$

for the defect system, where $f(\varepsilon) = (1 + e^{\varepsilon/(k_B T)})^{-1}$ and the temperature T is fixed. The existence theory of (31) using a variational method was performed in [Nie93]. For the defect problem (32), I proved the following theorem:

THEOREM III.3.1. *Fix a solution $V_{\rm per} \in L^2_{\rm per}$, $\varepsilon_F \in \mathbb{R}$ of (31). When $\mu_{\rm def}$ is small in $H^{-2}(\mathbb{R}^3)$, there is a unique solution $\rho_{\rm def}$ of (32) in a neighborhood of 0 in $L^2(\mathbb{R}^3)$. Furthermore, if $(1+|x|^2)^{1/2}\mu_{\rm def}$ is small in $H^{-2}(\mathbb{R}^3)$, then, for all $N \geq 1$, if $(1+|x|^2)^{N/2}\mu_{\rm def} \in L^2(\mathbb{R}^3)$, then $(1+|x|^2)^{N/2}V \in L^2(\mathbb{R}^3)$.*

This theorem applies to the defect potentials considered in [CL10], as well as to the case where $\mu_{\rm def} = Q\delta_0$, for Q small enough. In the latter case, it says that the total potential decays to zero at infinity faster than any polynomial (complete screening).

In the finite temperature case, the equation (29) for $\chi_{0,q}$ becomes

$$\chi_{0,q} = \frac{1}{|\mathcal{B}|}\int_{\mathcal{B}} \sum_{n,m\in\mathbb{N}} \frac{f(\varepsilon_{n,k+q} - \varepsilon_F) - f(\varepsilon_{mk} - \varepsilon_F)}{\varepsilon_{n,k+q} - \varepsilon_{mk}} |\overline{u_{mk}} u_{n,k+q}\rangle\langle \overline{u_{mk}} u_{n,k+q}|. \tag{33}$$

At finite temperature, this behaves like a metal: $(\chi_{0,q})_{00}$ tends to a non-zero limit as $q \to 0$, and therefore there is no divergence in $(A_q)_{00}$. Furthermore, and in contrast to the zero-temperature case, $\chi_{0,q}$ is a smooth function of q. This makes it possible to establish the decay properties of the linear response potential $A\mu_{\rm def}$, and, by a fixed-point argument, of V.

III.3.4. Charge sloshing. The above analysis is also relevant to understand the properties of algorithms to solve the self-consistent problem. We consider for simplicity the case of the potential mixing algorithm for the defect problem. Let G be the independent-particle mapping

$$G(V) = \Big(f(-\Delta + V_{\rm per} + V - \varepsilon_F) - f(-\Delta + V_{\rm per} - \varepsilon_F)\Big)(x,x).$$

from defect potential to defect electronic density. The mapping G is well-defined from a neighborhood of 0 in $L^2(\mathbb{R}^3)$ to $L^2(\mathbb{R}^3)$ (this results from a contour representation of $G(V)$ together with the Kato-Seiler-Simon inequality). Its differential at zero, χ_0, is a bounded

operator from $L^2(\mathbb{R}^3)$ to itself. Let $v_c = 4\pi(-\Delta)^{-1}$ be the (unbounded) Coulomb operator. Then the self-consistent equations (32) can be reformulated as

$$V = V_{\text{def}} + v_c G(V) \tag{34}$$

where $V_{\text{def}} = -v_c \mu_{\text{def}}$. The potential mixing algorithm is

$$V_{n+1} = V_n + \alpha(V_{\text{def}} + v_c G(V_n) - V_n).$$

Linearized around zero, this gives the Jacobian

$$J = 1 - \alpha(1 - v_c\chi_0).$$

Let us consider a truncation of space to a cubic domain of length L. Then the operator v_c, which acts as a multiplication in Fourier space by $1/|q|^2$, has eigenvalues proportional to L^2. In the case of zero-temperature insulators, as explained above, this is compensated by χ_0, so that $v_c\chi_0$ is a bounded operator; at finite temperature however, $v_c\chi_0$ has eigenvalues on the order of L^2. This means that α has to be chosen very small for the algorithm to converge. This appears in computations as charge moving along large-wavelength modes (for which v_c is large), called *charge sloshing* in the literature. This is however easy to remedy by using an appropriate preconditioner. In practice, for metallic systems, this is usually chosen as the Kerker scheme [Ker81], a multiplication operator in Fourier space given by

$$\mathcal{K}(q) = \frac{|q|^2}{\beta + |q|^2}$$

where β is a well-chosen constant. This acts as a high-pass filter which compensates for the divergence in v_c. I prove in **[2]** the following theorem:

THEOREM III.3.2. *Under the assumptions of Theorem III.3.1, the iteration*

$$V_{n+1} = V_n + \alpha\mathcal{K}(V_{\text{def}} + v_c G(V_n) - V_n)$$

converges for $\alpha > 0$ small enough.

The proof is an application of the Banach fixed point combined with the estimation that the spectrum of $\mathcal{K}(1 - v_c\chi_0)$ is bounded and strictly positive. This justifies the Kerker preconditioning scheme in the general setting of periodic metals (not limited to the homogeneous electron gas).

This study is limited to the finite-temperature case. The zero-temperature case involves additional technical complexities for screening, due to the appearance of Friedel oscillations; however, the analysis of charge sloshing should be similar. It is also restricted to periodic materials; it is known in practice that the self-consistent field iterations are slow to converge for inhomogeneous systems, and finding an adequate preconditioner is a challenge [LY13]. I intend to explore this in future work.

III.4. Independent electrons under a uniform electric field

In **[1]**, we considered another aspect of the response of periodic systems to external perturbations: electronic transport. Phenomenologically, when a uniform electric field is applied to a solid, one observes a flow of electrons in metals, and almost no current in insulators. Microscopically, this is again because in insulators, the electrons are tightly bound to their nuclei, and are not free to move: under the action of the electric field,

the electrons merely polarize, shifting their position slightly. In metals, by contrast, the classical Drude model assumes free electrons, which get accelerated by the electric field. This is ballistic transport, where the velocity of the electrons increases linearly. Then, after travelling in the crystal for some time τ, the electrons encounter an obstacle (a defect in the lattice, for instance), from which they bounce off at a random velocity. This ultimately results in a constant current flowing through the material, proportional to the imposed electric field (Ohm's law).

More exotic physics can be observed in regimes where quantum effects are important. In particular, we mention

- The quantum Hall effect, in which the transverse conductivity under a magnetic field is quantized in a very robust way;
- Bloch oscillations, in which electrons undergo oscillatory motion even without obstacles;
- The electronic properties of graphene, intermediate between insulators and metals.

III.4.1. Results. In [1], we investigated a very simple quantum model of non-interacting electrons in a perfect lattice that is able to represent these three effects. In dimension $d \leq 3$, let $V \in L^2_{\mathrm{per}}(\mathbb{R}^d, \mathbb{R})$ be a periodic scalar potential, and $\mathcal{A} \in L^4_{\mathrm{per}}(\mathbb{R}^d, \mathbb{R}^d)$ be a periodic vector potential. The unperturbed Hamiltonian is

$$H^0 = -\frac{1}{2}(-i\nabla + \mathcal{A})^2 + V,$$

a periodic operator with fibers

$$H^0_k = -\frac{1}{2}(-i\nabla + k + \mathcal{A})^2 + V.$$

We use the same notation as in the previous sections for its eigenvalues ε_{nk} and eigenvectors u_{nk}. The ground state of the electrons is modeled by the density matrix

$$\gamma(0) = \mathbb{1}(H^0 \leq \varepsilon_F), \tag{35}$$

where ε_F is a fixed Fermi level. Then, at time $t = 0$, we switch on a small uniform electric field in a fixed direction e_β:

$$H^\delta = H^0 + \delta x \cdot e_\beta$$

where $\delta \ll 1$, and solve the Liouville equation

$$i\frac{d\gamma^\delta}{dt}(t) = [H^\delta, \gamma^\delta(t)], \tag{36}$$

which is simply a reformulation of the time-dependent Schrödinger equation in the case of independent electrons.

In particular, we want to look at the evolution of the current per unit cell

$$j^\delta(t) = \underline{\mathrm{Tr}}(J_\alpha \gamma^\delta(t)),$$

in a fixed direction e_α, where $J_\alpha = -(-i\nabla + \mathcal{A}) \cdot e_\alpha$.

We say that a system is

- An insulator if there is $N \in \mathbb{N}$ such that

$$\sup_{k\in\mathcal{B}} \varepsilon_{Nk} < \varepsilon_F < \inf_{k\in\mathcal{B}} \varepsilon_{N+1,k}.$$

- A non-degenerate metal if $\varepsilon_F \in \sigma(H^0)$ and Assumptions III.2.1 and III.2.2 are satisfied (no crossing or flat bands at the Fermi level).
- A semi-metal if the dimension d is 2, the internal magnetic field $\mathcal{A}$ vanishes, and the Fermi surface is composed of a finite number N_{cross} of conical crossings $(k_i)_{i=1,\dots,N_{\mathrm{cross}}}$ such that

$$\varepsilon_{N_{\mathrm{sm}}-1,k_i} < \varepsilon_{N_{\mathrm{sm}},k_i} = \varepsilon_F = \varepsilon_{N_{\mathrm{sm}}+1,k_i} < \varepsilon_{N_{\mathrm{sm}}+2,k_i}$$

for some $N_i \in \mathbb{N}$, and that furthermore near k_i we have

$$\begin{aligned}\varepsilon_{N_{\mathrm{sm}},k} &= \varepsilon_F - v_{F,i}|k-k_i| + O(|k-k_i|^2)\\ \varepsilon_{N_{\mathrm{sm}}+1,k} &= \varepsilon_F + v_{F,i}|k-k_i| + O(|k-k_i|^2)\end{aligned}$$

for some Fermi velocity $v_{F,i} \neq 0$ (with the convention that $\varepsilon_{0,k} = -\infty$).

The assumption on the non-degenerate metals is the same as the one we have used in Section III.2. The assumption on the semi-metals is generic in the case of two-dimensional systems with honeycomb lattices, such as graphene [FW12].

We prove in [1] the following theorem.

THEOREM III.4.1. *For all $\delta > 0$, equations (35), (36) admit a unique solution $\gamma^\delta(t)$ for all times $t \in \mathbb{R}$, which is a bounded and periodic operator. The operator $J_\alpha \gamma^\delta(t)$ is locally trace class for all $t \in \mathbb{R}$, and the function j^δ is continuous.*

(a) *Assume that the system is an insulator. Then*

$$\lim_{t\to+\infty} \lim_{\delta\to 0} \frac{1}{\delta t} \int_0^t j^\delta(t')dt' = \frac{-i}{|\mathcal{B}|} \int_{\mathcal{B}} \mathrm{Tr}\left(\gamma_k(0)[\partial_{k_\alpha}\gamma_k(0), \partial_{k_\beta}\gamma_k(0)]\right) dk \tag{37}$$

(b) *Assume that the system is a non-degenerate metal. Then*

$$\lim_{t\to+\infty} \lim_{\delta\to 0} \frac{j^\delta(t)}{\delta t} = \frac{1}{|\mathcal{B}|} \sum_{n\in\mathbb{N}^*} \int_{S_n} (\nabla\varepsilon_{n,k} \cdot e_\alpha)\, (ds \cdot \vec{e}_\beta). \tag{38}$$

Furthermore assume that the Fermi level is crossed by an isolated band, i.e. that there exists $N \in \mathbb{N}^*$ such that $\varepsilon_{N-1,k} < \varepsilon_F < \varepsilon_{N+1,k}$ for all $k \in \mathcal{B}$ with uniform gaps between $\varepsilon_{N-1,k}$ and $\varepsilon_{N,k}$ on the one hand, and $\varepsilon_{N,k}$ and $\varepsilon_{N+1,k}$ on the other hand. Then there is $\eta > 0$ such that, for all $\delta, t \in \mathbb{R}$,

$$j^\delta(t) = \frac{1}{|\mathcal{B}|} \int_{\mathcal{B}} \mathbb{1}(\varepsilon_{N,k} \le \varepsilon_F) \partial_{k_\alpha} \varepsilon_{N,k-\delta e_\beta t} dk + O((\delta + \delta^2 t)\mathrm{e}^{\eta|t|\delta}). \tag{39}$$

(c) *Assume that the system is a semi-metal with N_{cross} conical crossings in the Brillouin zone, and that $V \in H^1_{\mathrm{per}}$. Then*

$$\lim_{t\to+\infty} \lim_{\delta\to 0} \frac{1}{\delta t} \int_0^t j^\delta(t')dt' = \frac{1}{|\mathcal{B}|} \frac{\pi^2}{4} N_{\mathrm{cross}}\, e_\alpha \cdot e_\beta. \tag{40}$$

III.4.2. Comments and numerical illustration. Our results show the following behavior, in their regimes of applicability:

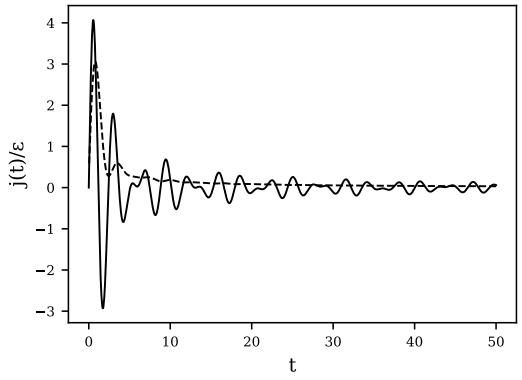
- In insulators, when the current is measured in the longitudinal direction ($e_\alpha = e_\beta$), the current vanishes when averaged in time. The transverse conductivity is related to the Chern numbers (see Section IV.4 in the next chapter) and is in particular quantized, a relationship known as the TKNN formula [TKNN82].
- In metals, in the regime $t \ll 1/\delta$, the current grows linearly with time, and electrons undergo ballistic transport. The proportionality factor, sometimes called Drude weight, is related to the density of charge carriers at the Fermi level.

  For larger times, due to a phenomenon known as Bloch oscillations, the current is actually oscillatory. For instance when e_β is a vector of the reciprocal lattice $\mathcal{R}^*$, the current is approximately periodic with period $1/\delta$. We are theoretically only able to observe a number of periods logarithmically growing with $1/\delta$. This results from our estimates involving an application of Gronwall's lemma to control the momentum; in a tight-binding model, we would have been able to observe a number of periods linearly growing with $1/\delta$. This condition may be related to the Ehrenfest time in semiclassical analysis [HJ00].
- In semimetals such as graphene, there is a finite conductivity, known as the residual conductivity. This conductivity is independent of the details of the crossings (and, in particular, of the Fermi velocities $v_{F,i}$).

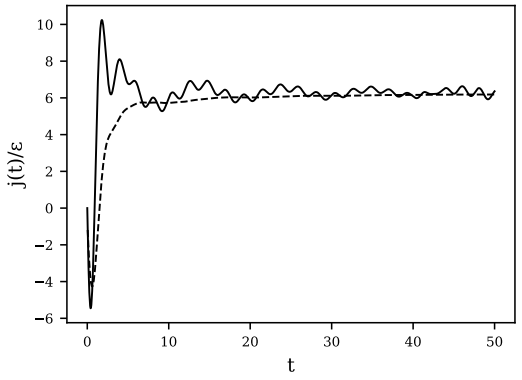
The situation can be visualized in the plots of Figure 4 in the regime $t \ll 1/\delta$ and Figure 5 for $t \gg 1/\delta$, where we refer to [**1**] for an explanation of the tight-binding 2D model used.

This study uses a very simple model of non-interacting electrons in a perfect lattice. In real materials, impurities and interactions with the motion of the nuclei affect the current. For insulators the effect is relatively minor: the conductivity, resulting from a topological property, is extremely robust to external perturbations, as is well-known from the quantum Hall effect. For metals, these effects act as obstacles to the ballistic propagation of the electrons, and eventually limit the value of the current and resulting in a finite conductivity (Ohm's law). A detailed mathematical investigation of this effect would be a very interesting direction of research.

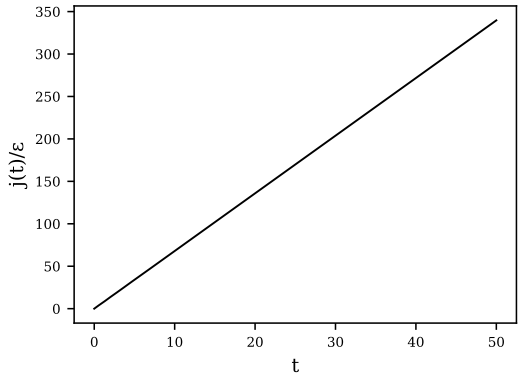
This and similar models have been studied by a number of authors in the mathematical community [PST03; BGKS05]. However, these studies have often considered insulators, using methods such as adiabatic switching or space-adiabatic perturbation theory to describe the equilibrium state of electrons in systems displaying the quantum Hall effect or Anderson localization. To our knowledge, ours is the first effort to describe periodic insulators, metals and semi-metals in an unified, mathematically rigorous, context.



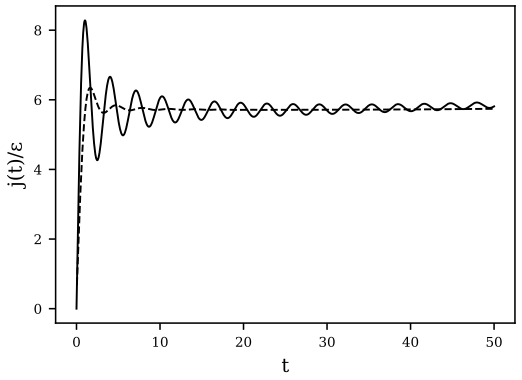
(A) Normal insulator phase.



(B) Chern insulator phase, transverse current.

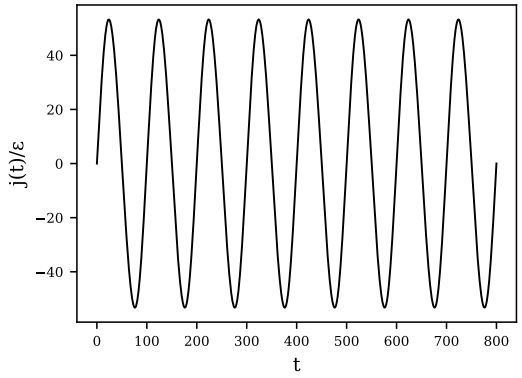


(C) Metallic phase.



(D) Semimetal.

FIGURE 4. Current $\frac{j^\delta(t)}{\delta}$ (solid line) and running average $\frac{1}{\delta t}\int_0^t j^\delta(t')dt'$ (dotted line) for several phases, in the linear response regime ($\delta = 10^{-6}$, $t \ll \frac{1}{\delta}$).



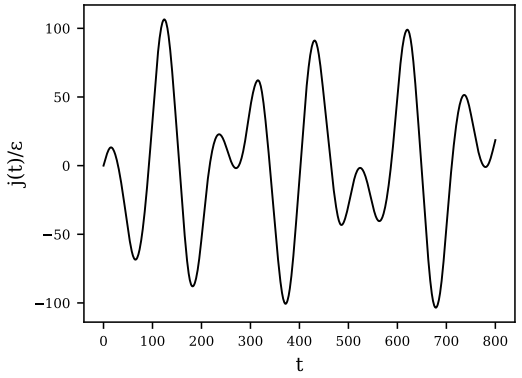


FIGURE 5. Current in the Bloch oscillations regime ($\delta = 10^{-2}$, $\frac{1}{\delta} \ll t$). The left panel is in a case where e_β is commensurate with the reciprocal lattice, the right panel is in a case where it is not.

III.4.3. Sketch of proof. We very briefly sketch the proof of Theorem III.4.1. First, it is surprising that $\gamma^\delta(t)$ is a periodic operator, even though H^δ is clearly not. The answer to this apparent paradox is that, while the potential δx_β is not periodic, its associated electric field $-\delta e_\beta$ is. To make this clearer, we perform the classical change of gauge

$$\widetilde{\gamma}(t) = e^{i\delta t x_\beta}\gamma(t)e^{-i\delta t x_\beta},$$

and obtain the equation

$$i\frac{d}{dt}\widetilde{\gamma}(t) = [\widetilde{H}^\delta(t), \widetilde{\gamma}(t)] \tag{41}$$

where

$$\widetilde{H}^\delta(t) = \frac{1}{2}(-i\nabla + \mathcal{A} - \delta e_\beta t)^2 + V$$

Therefore, we have turned a time-independent linear potential into a time-dependent uniform vector potential; this is consistent with the classical electromagnetic equation

$$E = -\nabla V - \frac{d\mathcal{A}}{dt}$$

for the electric field in a potential $(V, \mathcal{A})$, which shows that a constant uniform electric field can be imposed either through a constant but non-uniform scalar potential V, or through a uniform but time-dependent vector potential $\mathcal{A}$. The study of (36) can then be reduced to that of (41). In particular, this is now a periodic equation, which by Bloch theory reduces to the study of the fibers $\widetilde{H}^\delta_k(t)$.

We then have to study the dynamics generated by the time-dependent Hamiltonian

$$h(\delta t) := (-i\nabla + \mathcal{A} + k - \delta e_\beta t)^2 + V = H^0_{k-\delta e_\beta t}$$

for all $k \in \mathcal{B}$. For a given $k \in \mathcal{B}$, define the propagator $U^\delta(t)$ by $U^\delta(0) = 1$ and

$$i\frac{d}{dt}U^\delta(t) = h(\delta t)U^\delta(t).$$

Two tools can be used to compute approximations of $U^\delta(t)$ for δ small. We sketch these tools, in the case of a finite-dimensional Hilbert space for simplicity.

The first tool is linear response, which considers $h(\delta t)$ as a small perturbation of $h(0)$: for a given $t > 0$, $h(\delta t) = h(0) + \delta t h'(0) + O(\delta^2)$. Then, using Duhamel's formula, we can expand the propagator $U^\delta(t)$ in a Dyson series:

$$U^\delta(t) = e^{-ith(0)} - i\delta\int_0^t e^{-i(t-t')h(0)}h'(0)t'e^{-it'h(0)}dt' + O(\delta^2).$$

This formula can be used to obtain an expansion of the current to first order in δ, at fixed t (the Kubo formula). It is limited to short times $t \ll \frac{1}{\delta}$, but does not require any gap assumption.

To observe Bloch oscillations, one needs to access larger times. We can use there the adiabatic theorem [Teu03], which considers $h(\delta t)$ as a slow deformation of a family of Hamiltonians. If $P(\delta t)$ is the projector on a gapped spectral subspace of $h(\delta t)$, then

$$U^\delta(t)P(0)U^\delta(t)^* = P(\delta t) + i\delta L^+(\delta t)P'(\delta t) - i\delta U^\delta(t)L^+(0)P'(0)U^\delta(t)^* + O(\delta^2 t), \tag{42}$$

where L^+ is the Liouvillian pseudo-inverse (see [1] for the definition). To order 0, U^δ preserves the gapped eigenstates: this results in the ballistic transport of metals. The second term of this equation is static, and results in the finite transverse conductivity in insulators. The third term is oscillatory, and disappears when averaged in time.

Both linear response and adiabatic theories enable the computation of first-order properties, and both (a) for insulators and the first part of (b) for metals can be proved indifferently using either of these tools. For the second part of (b) (Bloch oscillations), only adiabatic theory is applicable, because of the need to access times of order $\frac{1}{\delta}$. In the case of semimetals, the presence of Dirac points that close the gap render adiabatic theory inapplicable, and only linear response can be used.

These arguments, together with a careful control on the remainder terms (in particular due to the unboundedness of the Hamiltonian), yield Theorem III.4.1 in the case of insulators and non-degenerate metals. For semimetals, a separate study is needed to control the divergence of oscillatory terms near the Fermi surface, which leads to the finite conductivity proved in Theorem III.4.1.

III.4.4. Step response of oscillatory systems. Finally, we comment on the order of limits, and on the time-averaging process. Our results take the limit as $\delta \to 0$, then as $t \to \infty$. The opposite limit $t \to \infty$ then $\delta \to 0$ is very hard to study. This is because our model is non-dissipative, and the long-time dynamics of non-dissipative systems (classical or quantum) are complicated out-of-equilibrium phenomena.

The limits do not in general commute. The simplest case where this can be seen explicitly is the case of a non-degenerate metal described by a tight-binding (or discrete Schrödinger) model, a point argued in [BESB94, Proposition 4]. In that case, when $e_\alpha = e_\beta$, $j^\delta(t)$ can be recognized as the derivative of a (bounded) energy, and so

$$\lim_{\delta\to 0}\lim_{t\to\infty}\frac{1}{\delta t}\int_0^t j^\delta(t')dt' = 0.$$

On the other hand, our results show that

$$\lim_{t\to\infty}\lim_{\delta\to 0}\frac{1}{\delta t}j^\delta(t) > 0.$$

Naively interpreted, these results show that the conductivity of metals is either 0 or infinity, depending on the definition. The reason why metals are observed to have a finite conductivity is due to the interaction of the electrons with obstacles such as impurities or nuclear motion, which is not taken into account in our simple model.

For insulators and semi-metals, our results apply not directly to the conductivity, but to its time-averaged value. This is because quantum dynamics is oscillatory, and these oscillations have to be taken out in one way or another. This can be illustrated on the very simple example of a driven linear oscillator

$$i\partial_t O(t) = \omega O(t) + I(t) \tag{43}$$

as would arise from the linear response of a two-level quantum system, where I (in our case, the electric field) is the input and O the output (in our case, the current per unit cell). The natural definition of the "conductivity" (response to a step function) in this system is to take $I(t) = 1$ and solve for the steady state $O(t) = -\frac{1}{\omega}$. But this steady state may never be reached. For instance, if $I(t)$ is turned to 1 brutally from 0 (as we

do in Theorem III.4.1), the solution of the equation for $t \geq 0$ is $O(t) = -\frac{1}{\omega}(1 - e^{-i\omega t})$, which oscillates and does not settle into the steady state. In our case, this appears as the oscillatory third term in (42).

This is a common problem, and has been dealt with in at least four different ways in the mathematical literature:

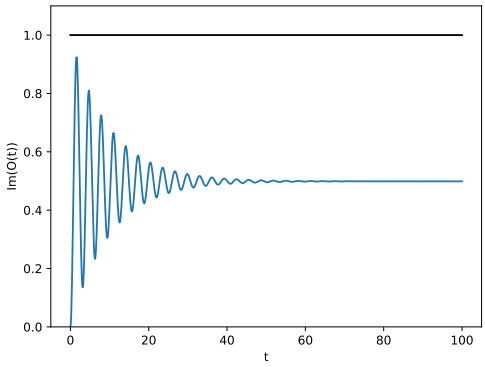
(1) Add a finite dissipation $\eta > 0$ to the system, for instance by $i\partial_t O_t + i\eta O(t) = \omega O(t) + I(t)$, compute the conductivity at finite η, and then let η tend to 0. This corresponds to the "relaxation time approximation" [BESB94] and is usually invoked to compute frequency-dependent response functions in computational physics.

(2) Define the conductivity through a time-averaging: $\sigma = \lim_{t\to\infty} \frac{1}{t} \int_0^t O(t')dt'$. This is the method we use in Theorem III.4.1.

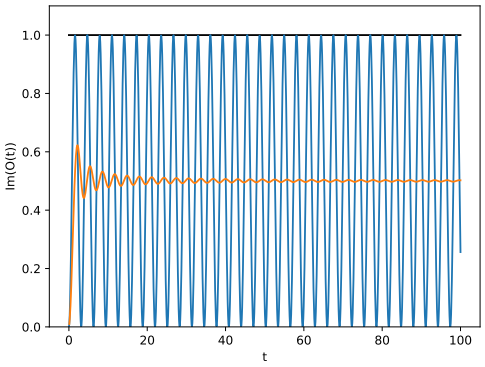
(3) Switch on the input I adiabatically, for instance by $I(t) = e^{\eta t}$ if $t < 0$, $I(t) = 1$ otherwise. This is for instance the approach used in [BGKS05].

For simple systems, it is easy to see that these methods are equivalent. This is illustrated in Figure 6.

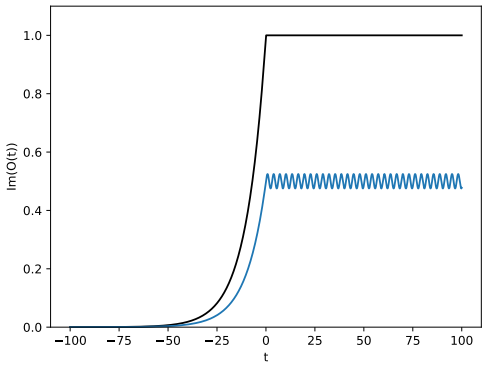
However, the numerical results in Figure 4 point to something more: the linear response current $\lim_{\delta\to 0} \frac{j^\delta(t)}{\delta}$ appears to tend to a finite limit as $t \to \infty$. This is not captured in our results, and results from a *dispersion* effect: unlike the simple model (43) that oscillates at a single frequency, extended systems possess a continuum of frequencies that interfere destructively for large times. Although this argument is not new either in mathematical physics or in the physical literature on electronic transport (it was already mentioned in the work of Kubo [Kub57]), to the best of our knowledge it has not been exploited mathematically in the context of macroscopic transport properties.



(A) Method 1, dissipation. The response decays as $\exp(-\eta t)$, and its eventual value is off by η.



(B) Method 2, time-averaging. The average converges as $1/(\omega t)$.



(C) Method 3, adiabatic switching. The oscillation size is proportional to η.

FIGURE 6. Three ways of computing the step response of an oscillatory system, presented on the equation $i\partial_t O = \omega O + I$. We use the settings $\omega = -2, \eta = 0.1$. The end result is the same in all methods: a conductivity of $-1/\omega$. The input I is in black, the output O in blue, the time-averaged output $\frac{1}{t}\int_0^t O(t')dt'$ in yellow.

CHAPTER IV

Wannier functions

Wannier functions, first introduced as a conceptual tool in 1937 [Wan37], are the closest analogue to localized orbitals that can be obtained in the crystalline phase. As we have seen in Chapter III, the spectrum of periodic Schrödinger operators is continuous, with Bloch waves

$$\psi_{nk}(x) = e^{ik\cdot x} u_{nk}(x)$$

as generalized eigenvectors, where $k \in \mathcal{B}$ and u_{nk} is $\mathcal{R}$-periodic. These Bloch waves are delocalized over all physical space. This is at odds with the physical picture of electrons in insulators staying close to their nuclei. Wannier functions are a way to reconcile these two pictures.

Note that this is simply a more extreme version of a problem already faced in molecules: generally, the eigenfunctions of the Hamiltonian of a large molecular system have orbitals $(\phi_n)_{n=1,\dots,N}$ delocalized over the whole molecule. In that case, an equivalent physical description can be obtained by rotating the orbitals:

$$\widetilde{\phi}_n(x) = \sum_{m=1}^{N} \phi_m(x) U_{mn}$$

for a given unitary matrix $U \in \mathcal{U}(N)$. For example, in the Hartree-Fock method, both sets ϕ_n and $\widetilde{\phi}_n$ of orbitals give rise to the same Slater determinant. By playing with the coefficients U_{mn}, one can localize the $\widetilde{\phi}_n$ near the atoms (or groups of atoms), and from there get a more directly interpretable set of orbitals [FB60].

Turning back to the periodic setting, consider an insulator with N occupied bands. We seek a unitary combination of the Bloch waves $(\psi_{nk})_{n=1,\dots,N,\, k\in\mathcal{B}}$ that are localized. In principle, one could take arbitrary combinations, of the form

$$\sum_{m=1}^{N} \alpha_m(k) u_{mk}(x) e^{ik\cdot x} dk$$

for arbitrary complex-valued functions $\alpha_m(k)$. This however would yield too rich a set, and would break periodicity in the set of localized orbitals. The idea of Wannier was to require functions to be translates of each other, i.e. to find an orthogonal set $(w_{nR})_{n=1,\dots,N,\, R\in\mathcal{R}}$ such that $w_{nR} = w_n(\cdot - R)$. It is easy to check that this implies the particular structure

$$w_{nR}(x) = \frac{1}{|\mathcal{B}|} \int_{\mathcal{B}} \sum_{m=1}^{N} u_{mk}(x) U_{mn}(k) e^{ik\cdot(x-R)} dk \tag{44}$$

where the orthogonality condition $\langle w_{nR}, w_{n'R'}\rangle = \delta_{nn'}\delta_{RR'}$ constrains the matrices $U(k)$ to be unitary. This equation defines the Wannier functions w_{nR}.

Note that the U_{mn} in this equation are free, as they were in the molecular case. The question is then: can we choose U_{mn} to localize the w_{nR}? This is intimately related to the smoothness in the Brillouin zone of the pseudo-Bloch functions

$$v_{nk}(x) = \sum_{m=1}^{N} u_{mk}(x) U_{mn}(k).$$

Indeed, if v_{nk} has the Fourier series expansion $v_{nk}(x) = \sum_{K\in\mathcal{R}^*} c_{nk}(K)e^{iK\cdot x}$, then w_{nR} has the expansion

$$w_{nR}(x) = \frac{1}{|\mathcal{B}|}\int_{\mathcal{B}} \sum_{K\in\mathcal{R}^*} c_{nk}(K)e^{i(k+K)\cdot(x-R)}dk.$$

from which it follows by standard Fourier duality that the locality of w_{nR} is linked with the smoothness of the map $k + K \mapsto c_{nk}(K)$. In particular, if $\mathbb{R}^3 \ni k \mapsto v_{nk}$ is infinitely differentiable and satisfies $v_{n,k+K}(x) = e^{-iK\cdot x}v_{nk}(x)$ for all $K \in \mathcal{R}^*$, then the corresponding Wannier functions decay faster than any inverse polynomial; if v_{nk} is analytic on a strip in the complex domain, then the Wannier functions decay exponentially. Therefore, finding localized Wannier functions is equivalent to finding a smooth v_{nk} in Bloch space. As we will see, this is linked to the numerical problem of interpolating spectral information in the Brillouin zone.

Localized Wannier functions, if they can be constructed, are a useful interpretative and numerical tool. They allow us to:

- Interpret the nature of chemical bonding in materials by plotting the w_{nR};
- Understand phenomena such as polarization in solids (the "modern theory of polarization" [RV07]);
- Explore the topological structure of materials;
- Reduce continuous models to discrete ones by expanding the Hamiltonian into the Wannier functions basis;
- Add additional correlation into a KSDFT model, for instance by the LDA+U method;
- Speed up computations such as Hartree-Fock exchange by exploiting locality;
- Enhance the sampling of the Brillouin zone by interpolation.

We refer to [MMYSV12] for an overview of these applications. The questions facing the applied mathematician are then

- Under what conditions does there exist localized Wannier functions?
- How localized are they?
- How can we compute them in practice?
- Can an analogous theory be formulated for metals?

The first two questions were adressed comprehensively by various authors, starting in the '60s and finally being settled in 2007, the central result being that exponentially-localized Wannier functions exist under the geometrical condition that the Chern numbers of the system under consideration vanish [Pan07; BPCMM07]. The third and fourth questions were adressed numerically by localization schemes in [MV97], without guarantee that the scheme produces localized Wannier functions. This scheme was extended to metals in [SMV01], but lacks a theoretical basis.

After reviewing the theory of Wannier functions and their relationship with the topological properties of materials, I will describe my contributions. In **[7]** and **[4]**, described in Sections IV.5 and IV.6, we propose a method for the construction of localized Wannier for insulators, including topological insulators. In **[6]**, described in Section IV.7, we extend the theory of Wannier function localization to the metallic case. In **[5]**, described in Section IV.8, we study optimization algorithms for their construction. In order to make the exposition self-contained and to illustrate its relevance in a broader applied mathematics

context, I will review the theory of Wannier functions from the point of view of eigenvalue interpolation.

IV.1. Eigenvalue interpolation

Recall that the spectrum of a self-adjoint locally compact-resolvent periodic operator H (for instance, the mean-field Hamiltonian arising from a DFT computation in a periodic system) can be obtained by solving the eigenvalue problem

$$H_k u_{nk} = \varepsilon_{nk} u_{nk}$$

for $u_{nk} \in L^2_{\mathrm{per}}$, for every k in the Brillouin zone $\mathcal{B}$, and where

$$H_k = e^{-ik\cdot x} H e^{ik\cdot x}.$$

Properties of interest are then expressed as integrals of quantities depending on the u_{nk} and ε_{nk} over the Brillouin zone $\mathbb{R}^3/\mathcal{R}^*$.

Computing the u_{nk} and ε_{nk} for a given k requires the solution of an eigenvalue problem, and is therefore expensive. It is common for quantities of interest (especially response properties of metals) to require thousands or millions of k-points to converge to a useful value. In practice however, one is interested only in a subset of the bands: those that are low in energy (for ground state properties), or close to the Fermi level (for response properties). It is therefore desirable to interpolate k-dependent values such as the ε_{nk}, the u_{nk} or derived quantities, in that energy window. For simplicity, we take that window to be the index set $I \subset \mathbb{N}$, assume that H_k has been discretized to a matrix by one of the methods described in Chapter II, and that we have computed the $(u_{nk})_{n\in I}$ and $(\varepsilon_{nk})_{n\in I}$ on a discrete grid in $\mathcal{B}$. We are now interested in interpolating the ε_{nk} in the full $\mathcal{B}$; other quantities of interest, such as the matrix elements of various operators in the u_{nk} basis, can be interpolated in a similar way [MMYSV12].

Note that, although H_k is not $\mathcal{R}^*$-periodic, the eigenvalues ε_{nk} are. This is because H_k satisfies

$$H_{k+K} = \tau_K H_k \tau_K^* \tag{45}$$

for every $K \in \mathcal{R}^*$, and where

$$(\tau_K u)(x) = e^{-iK\cdot x} u(x)$$

is a unitary operator on L^2_{per}. The operators H_{k+K} and H_k therefore share the same eigenvalues [1]. We will say that a family of operators is τ-periodic if it satisfies (45), and a family of vectors $\mathbb{R}^3 \ni k \mapsto u_k$ is τ-periodic if $u_{k+K} = \tau_K u_k$ for all $K \in \mathcal{R}^*$. Our problem is then:

Problem IV.1.1. *Given a smooth τ-periodic family of matrices H_k with eigenvalues ε_{nk} and an index set $I \subset \mathbb{N}$, how can we efficiently interpolate $(\varepsilon_{nk})_{n\in I}$ in $\mathcal{B}$?*

[1] Note that this condition is not exactly satisfied by the matrices resulting from a discretization of H_k. A fixed basis for every k cannot ensure this property, because no non-trivial finite-dimensional basis is left invariant by τ_K. Plane-wave discretization usually have a $k-$dependent basis that includes all wave vectors $G \in \mathcal{R}^*$ such that $\frac{1}{2}|k+G|^2 \leq E_{\mathrm{cut}}$: the matrix H_k is then τ-periodic but not smooth, as the basis changes discontinuously. Although an analysis of this effect would be interesting, it is usually not a problem in practice for pseudopotential methods as the convergence with respect to E_{cut} is rapid, and so we neglect it in the following.

The main obtacle to the direct interpolation of the ε_{nk} is eigenvalue crossings, at which the eigenvalues lose their smoothness. A trivial example is the matrix

$$H(k) = \begin{pmatrix} \cos(2\pi k) & 0 \\ 0 & -\cos(2\pi k) \end{pmatrix},$$

periodic with period 1 and $\tau = 1$, and with eigenvalues

$$\varepsilon^{\pm}(k) = \pm|\cos(2\pi k)|$$

Attempting to interpolate these two eigenvalues directly with high-order interpolation schemes such as Fourier interpolation results in the Gibbs phenomenon: the slowly-decreasing Fourier components of $\varepsilon^{\pm}$ cause the interpolation to converge slowly, as seen Figure 1.

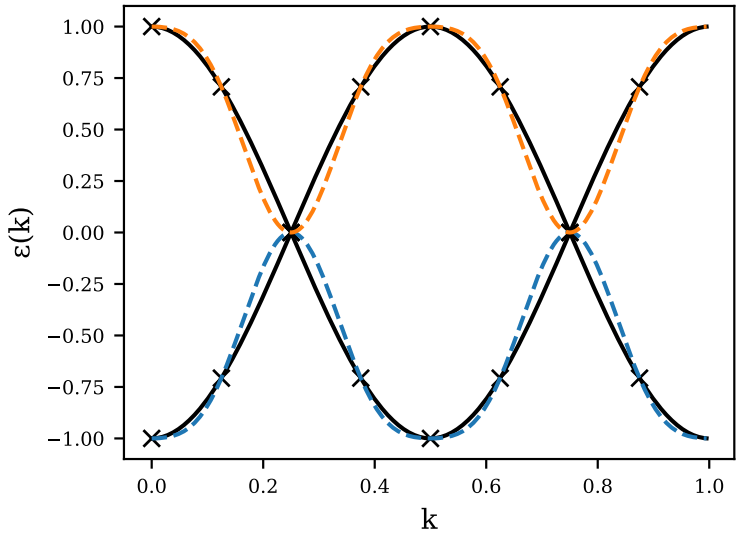


FIGURE 1. Eigenvalues $\varepsilon^{\pm}(k) = \pm|\cos(2\pi k)|$ (plain), and their Fourier interpolation (dashed) at $N = 8$ data points in $[0, 1]$. A Gibbs phenomenon is present because of the eigenvalue crossings at $\frac{1}{4}$ and $\frac{3}{4}$.

On this problem, the fix is trivial: connect the bands, and interpolate instead the functions $\pm\cos(2\pi k)$. However, this does not work in higher dimensions, where for instance the eigenvalues

$$\varepsilon_{\pm} = \pm\sqrt{\sin(2\pi k_1)^2 + \sin(2\pi k_2)^2}$$

of

$$H(k) = \begin{pmatrix} \sin(2\pi k_1) & \sin(2\pi k_2) \\ \sin(2\pi k_2) & -\sin(2\pi k_1) \end{pmatrix}$$

cannot be relabelled in a smooth way around the conical intersection at $(0, 0)$.

The problem we described comes from the fact the map from a matrix to its eigenvalues is not smooth at eigenvalue crossings. This suggests bypassing this operation and interpolating $H(k)$ directly: interpolate-and-diagonalize rather than diagonalize-and-interpolate. However, H is generally too large for this procedure to work. We therefore seek a reduced Hamiltonian $\widetilde{H}(k)$ whose eigenvalues are those of $H(k)$ in a given energy window $I \subset \mathbb{N}$. Interpolating the u_{nk} directly encounters two obstacles. First, the u_{nk} are complex eigenvectors: they are only defined up to a phase, and therefore have no reason to be continuous from one k-point to another. Second, at eigenvalue crossings the u_{nk} are discontinuous, even with a proper choice of phase. However, if one can find a set $(v_{nk})_{n\in I}$ that spans $X(k)$ and is smooth as a function of $k \in \mathbb{R}^d$, then one can form the reduced matrix

$$\widetilde{H}_{mn}(k) = \langle v_{mk}, H_k v_{nk} \rangle,$$

and interpolate its components by Fourier interpolation. The values of ε_{nk} and other quantities of interest on off-grid points can then be recovered by a diagonalization of the interpolated small-size matrix $\widetilde{H}(k)$.

Note that for this to be feasible, we need that the subspace $X(k)$ spanned by the $(u_{nk})_{n\in I}$ is smooth as a function of k, which is only possible if the eigenvalues $(\varepsilon_{nk})_{n\in I}$ are separated from the rest of the spectrum $(\varepsilon_{nk})_{n\notin I}$:

$$\inf_{k\in\mathcal{B}, i\in I, a\notin I} |\varepsilon_{ik} - \varepsilon_{ak}| > 0,$$

which we assume in the following. This is the case for instance in insulators, where $I = \{1, \dots, N\}$ with N the number of electrons per unit cell.

Our problem can then be restated as

Problem IV.1.2. *Given a smooth τ-periodic family of subspaces $X(k)$, find a smooth and τ-periodic basis v_{nk} of $X(k)$.*

IV.2. Localizing Wannier functions

As we have seen, solving Problem IV.1.2 is equivalent to finding localized Wannier functions

$$w_{nR}(x) = \frac{1}{|\mathcal{B}|} \int_{\mathcal{B}} v_{nk}(x) e^{ik(x-R)} dk.$$

It is therefore an important numerical problem to obtain well-localized Wannier functions. This program was started in 1997 by the introduction of maximally-localized Wannier functions (MLWF) by Marzari and Vanderbilt [MV97]. This numerical procedure minimizes iteratively the total variance

$$\Omega = \sum_{n\in I} \left(\int_{\mathbb{R}^3} |x|^2 |w_{n0}|^2(x) dx - \left| \int_{\mathbb{R}^3} x |w_{n0}|^2(x) dx \right|^2 \right), \tag{46}$$

a measure of localization. Again by Fourier duality, this can be seen as measuring smoothness in k-space, and is analogous to a Dirichlet energy in that space. Picking a set of localized Wannier functions with small Ω (equivalently, a smooth basis v_{nk}) will then ensure an accurate interpolation scheme.

Due to the non-convexity of the orthogonality constraints, minimizing this among all the possible Wannier functions (equivalently, among the possible orthogonal bases v_{nk}) is

a highly non-convex optimization problem, whose landscape is plagued by the presence of topological vortices, as we will see later. It is therefore crucial to start with a good initial point. This is usually done in practice by the projection method. One starts with an initial guess of the functions w_{n0} from chemical intuition of the system and bands under consideration (for instance, bond-centered s-like orbitals in silicon, or atom-centered p_z-like orbitals in graphene), projects those functions onto the spectral subspace of interest, and then renormalizes them [MV97]. These are then used as initial guess in an iterative minimization of the total variance Ω. This yields well-localized Wannier functions when the initial guess is good enough; when it is not, it yields spurious local minima that are not localized.

Recently, the field of computational materials science has moved towards the automated computation of material properties in large-scale databases (see for instance [JOHCR+13; PCSMK16]). In that context, hand-picking Wannier functions by trial-and-error of various initial guesses is not feasible. It is then desirable to design an algorithm that does not rely on chemical intuition for an initial guess, and is able to produce reliably well-localized Wannier functions, without human input. Work in that direction has been undertaken by [DLY15], which uses a rank-revealing factorization of the density matrix, and in [MCCL15], starting from an extended set of projections. However, these methods either require system-specific information, or fail in certain specific cases. In the following, I will describe our fully robust approach, based on an implementation of a constructive answer to Problem IV.1.2.

For the sake of exposition, I will make a number of simplifying assumptions in the remainder of this section

- The system is two-dimensional (the extension to the 3D case does not pose significant additional complications);
- We take $\mathcal{R}^* = \mathbb{Z}^2$ and $\tau = 1$ (this restriction can be lifted by a change of variable and appropriate treatment of boundary);
- The underlying Hilbert space is finite-dimensional (the construction can be easily generalized);
- We only construct continuous bases; such bases can be post-processed to be smooth, theoretically by the arguments of [FMP16, Section 5], and numerically by feeding them to the Marzari-Vanderbilt optimization procedure.

We refer to [7] and [4] for further details.

IV.3. Parallel transport

If we drop the condition of periodicity, Problem IV.1.2 can be solved using the notion of parallel transport. Parallel transport refers to a procedure for transporting a basis of a subspace when that subspace changes. If $\mathbb{R} \ni t \mapsto X(t)$ is a smooth subspace and $v(0)$ an orthogonal basis of $X(0)$, parallel transport constructs an orthogonal basis $v(t)$ of $X(t)$ through the solution of the following differential equation:

$$\frac{dv}{dt}(t) = \frac{dP_{X(t)}}{dt} v(t)$$

where P_X is the projector on the subspace X. This is analogous to the parallel transport of tangent vectors along a curve on a manifold. It is well known that this transport yields

information on the curvature on the underlying manifold. A simple visualization in the case of the sphere is given in Figure 2. When moving along the $A \to N \to B \to A$ path, the vector acquires an angle α. For infinitesimally small paths, this angle is proportional to the curvature. Here, in a similar way, if $X(t)$ describes a circular path with $X(1) = X(0)$, then $v(1) \neq v(0)$, the difference being related to the (Berry) curvature of $X(t)$ [Ber84; Sim83].

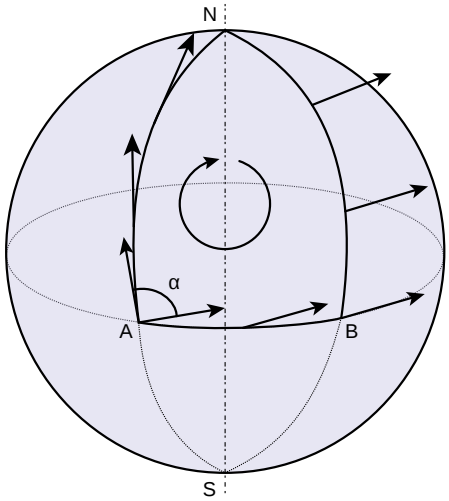


FIGURE 2. Parallel transport on the sphere. Source: https://en.wikipedia.org/wiki/File:Parallel_Transport.svg (Creative Commons license).

Using parallel transport, one can solve Problem IV.1.2 without the periodicity requirement. Start from an arbitrary $v(0,0)$. Then, solve the following differential equation

$$\frac{dv}{dk_1}(k_1,0) = \frac{dP_{X(k_1,0)}}{dk_1} v(k_1,0)$$

with initial condition $v(0,0)$ to obtain a smooth basis $[0,1] \ni k_1 \mapsto v(k_1,0)$ of $X(k_1,0)$. Then, for all $k_1 \in [0,1]$, solve

$$\frac{dv}{dk_2}(k_1,k_2) = \frac{dP_{X(k_1,k_2)}}{dk_2} v(k_1,k_2)$$

with initial condition $v(k_1,0)$ to obtain a smooth basis $[0,1]^2 \ni k \mapsto v(k)$ of $X(k)$. The problem is now that the v obtained in this fashion is not periodic: even in one dimension, $v(1) \neq v(0)$.

It is instructive to test this in one dimension on the matrix

$$H(k) = \begin{pmatrix} -\cos(2\pi k) & \sin(2\pi k) \\ \sin(2\pi k) & \cos(2\pi k) \end{pmatrix}$$

with $\mathcal{B} = [0,1]$ and $X(k)$ the eigenspace associated with the lowest eigenvalue $\varepsilon_{1k} = -1$ of $H(k)$. When run with $v(0) = (1,0)$, it gives $v(k) = (\cos(\pi k), \sin(\pi k))$. This satisfies $v(1) = -v(0)$. The geometrical situation is that of a Möbius strip: when making a turn

around the strip (k from 0 to 1), the orientation of the strip (the sign of v) gets reversed. This "geometric phase" has important consequences for quantum dynamics around conical intersections [JDRI13]. In this situation, there is no way to solve Problem IV.1.2 while keeping v real; however, we can choose $v(k) = e^{i\pi k}(\cos(\pi k), \sin(\pi k))$, undoing the sign change with a phase rotation.

This example can be generalized to show that Problem IV.1.2 can always be solved in 1D: start from an arbitrary $v(0)$ and use parallel transport to extend it to $\widetilde{v}(k)$ for $k \in [0, 1]$. Then, we have $\widetilde{v}(1) \neq \widetilde{v}(0)$ in general, but they both span the same space, and so there is a unitary matrix U such that

$$\widetilde{v}(1) = \widetilde{v}(0)U.$$

Then, define $v(k)$ by

$$v(k) = \widetilde{v}(k)U^{-k},$$

where $U^k = \exp(k \log U)$ and we used the principal branch for the logarithm of complex numbers: $\operatorname{Im} \log z \in (-\pi, \pi]$ for all $z \in \mathbb{C}^*$. This $v(k)$ is continuous, and satisfies $v(1) = v(0)$, hence solving the problem.

IV.4. Topology and Chern number

However, difficulties occur in dimension 2 and 3. These are characterized by the following classical theorem:

THEOREM IV.4.1. *Let S be a smooth, compact, oriented two-dimensional surface, and $S \ni k \mapsto X(k)$ be a smooth family of complex finite-dimensional spaces. Then there exists a continuous orthogonal basis $v(k)$ of $X(k)$ if and only if $\operatorname{Ch}(S, X) = 0$, where the Chern number is given by*

$$\operatorname{Ch}(S, X) := \frac{1}{2\pi i} \int_S \operatorname{Tr}(P_X dP_X \wedge dP_X) \tag{47}$$

with $P_X(k)$ the projector on $X(k)$.

The 2-form $-i \operatorname{Tr}(P_X dP_X \wedge dP_X)$ is known as the Berry curvature. When S is the 2-torus, identified to $[0, 1]^2$ with periodic boundary conditions, this formula simplifies to

$$\operatorname{Ch}(X) = \frac{1}{2\pi i} \int_{[0,1]^2} \operatorname{Tr}(P_{X_k}[\partial_{k_1} P_{X_k}, \partial_{k_2} P_{X_k}]).$$

(compare with (37) in the previous chapter).

This theorem is classical; see **[6]** Lemma 3.2 for a self-contained proof. The quantity $\operatorname{Ch}(S, X)$ is an integer, known as the Chern number. This is a topological quantity: since it depends continuously on X and is an integer, it will be the same for all small perturbations of X.

The prototypical example of a non-zero Chern number is the subspace associated with the first eigenvalue -1 of the Hamiltonian

$$H(k) = k \cdot \sigma = \begin{pmatrix} k_3 & k_1 - ik_2 \\ k_1 + ik_2 & -k_3 \end{pmatrix}$$

on the sphere $S = \{k \in \mathbb{R}^3, |k| = 1\}$. A particular choice of eigenvector is given by

$$\begin{pmatrix} \cos\left(\frac{\theta}{2}\right) e^{-i\phi} \\ \sin\left(\frac{\theta}{2}\right) \end{pmatrix}$$

in spherical coordinates $k = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$. At $\theta = 0$ (the north pole), this is not well-defined (because ϕ is not defined there). In a small circle of latitude near the north pole, the eigenvector above makes one full turn. One can attempt to remedy this singularity by introducing a global change of phase, but this will only succeed in moving the singularity away from the north pole. In fact, the total signed number of singularities (counted according to the oriented number of turns of the phase near the singularity) is a topological invariant, which is exactly what is measured by the Chern number: in this case, we have $\mathrm{Ch}(S, X) = 1$.

This example was on the sphere, but can easily be mapped on the torus (for instance through a map that sends the boundary of a square to the south pole of the sphere, and the center to the north pole). Therefore, to a given periodic physical system and isolated set of bands we can associate a non-trivial integer. Far from being a mathematical curiosity, this integer is physically measurable as a transport property (see Section III.4), and is responsible for the quantization of the conductivity in the (integer) quantum Hall effect [TKNN82].

IV.5. Finding smooth and periodic bases numerically

We see from the above Theorem IV.4.1 that, given an isolated set of bands in a two-dimensional system, one can solve Problem IV.1.2 to get an interpolation for the spectral information of these bands if and only if the associated Chern number is zero. This is the case in systems without magnetic fields, modeled by Hamiltonians of the form $H = (-i\nabla + k)^2 + V$. We have $H_{-k} = \overline{H_k}$ and

$$P_{X_{-k}} = \overline{P_{X_k}}, \tag{48}$$

which is a manifestation of time-reversal symmetry. In this case, it can be checked that the integrand in the equation (47) defining the Chern number is odd [Pan07]. Accordingly, the Chern number vanishes and it is always possible to solve Problem IV.1.2. This condition (48) is broken in the presence of a magnetic field, whether imposed externally (as in the case of the quantum Hall effect [TKNN82]) or by the material itself (as in the case of the quantum anomalous Hall effect [Hal88]).

In [**7**] and [**4**], we propose a numerical method to solve Problem IV.1.2 constructively, and test it on several systems of interest. Our method proceeds by mimicking the proof of Theorem IV.4.1. We describe the method in two dimensions, the extension to the three-dimensional case being very similar.

First, we build a continuous basis $v(k, 0)$ of $X(k, 0)$ on $[0, 1] \times \{0\}$ by using the above construction: we pick a basis $\widetilde{v}(0, 0)$ of $X(0, 0)$ at $(0, 0)$, propagate it by parallel transport to $\widetilde{v}(k, 0)$ along the segment $[0, 1] \times \{0\}$, find the unitary matrix U such that $\widetilde{v}(1, 0) = \widetilde{v}(0, 0)U$, and set $v(k, 0) = \widetilde{v}(k, 0)U^{-k}$. Then, for every $k_1 \in [0, 1]$, we propagate $v(k_1, 0)$ using parallel transport in the k_2 direction to $v(k_1, k_2)$. This $v(k_1, k_2)$ is periodic in the

k_1 direction, but not in the k_2 direction: we have

$$v(k,1) = v(k,0)U(k) \tag{49}$$

where U is a periodic family of unitary matrices.

This is where possible topological obstructions appear. We recall that, if $k \mapsto z(k)$ is a 1-periodic family of complex numbers of modulus one, the winding number is an integer defined through

$$W(z) = \frac{1}{2\pi i}\int_0^1 \frac{z'(k)}{z(k)}dk.$$

Linking $U(k)$ with the Berry connection and using an integration by parts argument [FMP16] shows that $\det U(k)$ is the Chern number associated with X_k on the torus $[0,1]^2$. It follows that, if one can find a smooth and periodic basis $w(k_1,k_2)$ of X_{k_1,k_2}, so that w satisfies (49) with $U = 1$, the Chern number must vanish.

To solve Problem IV.1.2, we are after the converse: provided the Chern number vanishes, can we find a smooth and periodic basis? We can then reduce this question to the following homotopy subproblem:

PROBLEM IV.5.1. *Given a family $k \mapsto U(k)$ of smooth and 1-periodic unitary matrices such that the winding number of $\det U$ vanishes, find a homotopy to the identity, i.e. a smooth family $U(k,t)$ of unitary matrices such that*

- $\forall k \in [0,1]$, $U(k,0) = 1$
- $\forall k \in [0,1]$, $U(k,1) = U(k)$
- $\forall t \in [0,1]$, $U(0,t) = U(1,t)$

Assuming that this problem can be solved, we can solve Problem IV.1.2 by constructing a continuous and periodic basis through

$$w(k_1,k_2) = v(k_1,k_2)U(k_1,k_2)^{-1}.$$

We focus on Problem IV.5.1 in the sequel.

In the case of a one-dimensional subspace ($N = 1$, so that $U(\cdot) \in \mathcal{U}(1)$), one can write $U(k) = e^{i\theta(k)}$, where the phase $\theta(k)$ can be chosen to be continuous on $[0,1]$. We have $W(U) = \frac{1}{2\pi}(\theta(1) - \theta(0))$, so that θ is periodic. We can then set

$$U(k,t) = e^{it\theta(k)}$$

to solve the problem. When $N > 1$, one can try to diagonalize the matrix $U(k)$ and use the previous construction eigenvector by eigenvector, i.e. set

$$U(k,t) = U(k)^t$$

where the branch cut of the logarithm is chosen per eigenvector to ensure continuity. The method we proposed in [7] did exactly that. However, this sometimes fails: for instance, if $U(k) = \mathrm{diag}(e^{2\pi ik}, e^{-2\pi ik})$, there are no continuous and periodic Hermitian matrices $L(k)$ such that $U(k) = e^{iL(k)}$. This is found in practice in materials presenting a strong spin-orbit interaction, which, while still preserving a time-reversal symmetry similar to (48), leads to a topological phase described by a $\mathbb{Z}_2$ topological index, such as the Kane-Mele model [KM05].

In [4], we revisited Problem IV.5.1 and provided an algorithm that, although more complex than that of [7], was guaranteed to always solve the problem, being based on a

constructive solution to Problem IV.5.1. In algebraic terms, the reasoning above says that the homotopy class of $\mathcal{U}(1)$ is $\mathbb{Z}$: any loop in $\mathcal{U}(1)$ can be deformed to any other loop, provided the winding number $W(U)$ is preserved. What then is the homotopy class of $\mathcal{U}(N)$? It turns out that it is also $\mathbb{Z}$: any loop in $\mathcal{U}(N)$ can be deformed to any other loop, provided that the winding number $W(\det U)$ is preserved. The proof of that fact usually proceeds through a fibration argument that is not easily translated into a constructive algorithm.

In [4], we proposed an algorithm based on column interpolation. Schematically, the construction is as follows (in practice, we use a slightly more complex and efficient construction, but the basic idea is similar). Given a loop $U(k)$ in $\mathcal{U}(N)$, its first column $u_1(k)$ describes a loop in the sphere $\{u \in \mathbb{C}^N, |u| = 1\}$. Since the sphere is simply connected, this loop can be deformed to a single point $e_1 = (1, 0 \cdots, 0)$. At the same time as $u_1(k)$ is deformed, we can deform by parallel transport on $\mathrm{Ran}(u_1(k))^\perp$ the other columns $u_2(k), \cdots, u_N(k)$, to preserve the orthogonality of $U(k)$. This establishes a deformation in $\mathcal{U}(N)$ from the loop $U(k)$ to a loop $\widetilde{U}(k)$, such that the first column of $\widetilde{U}(k)$ is e_1. It follows from unitarity that $\widetilde{U}(k)$ is of the form

$$\widetilde{U}(k) = \begin{pmatrix} 1 & 0 \\ 0 & V(k) \end{pmatrix}$$

Furthermore, $W(V) = W(\widetilde{U}) = W(U) = 0$. We can then proceed by induction over N.

IV.6. Implementation and results

In practice, we implement the algorithm above by discretizing the Brillouin zone $[0, 1]^2$ with a uniformly spaced grid. At each grid point k, we compute the eigenvectors $(u_n(k))_{1 \leq n \leq N}$ of H_k, with arbitrary phase. We now look for a unitary matrix $U_{mn}(k)$ such that $\sum_{m=1}^{N} u_m(k) U_{mn}(k)$ varies smoothly as a function of k. We apply the above algorithm using a discretization of parallel transport (see [7]). For the homotopy method, we use a variant of the algorithm described above that deforms the first column $u_1(k)$ to a reference point $\underline{u}$ instead of e_1. We determine $\underline{u}$ to ensure that $u_1(k)$ is well-separated from $\underline{u}$ for all k, and then use the interpolation

$$u_1(k, t) = \frac{(1-t)u_1(k) + t\underline{u}}{|(1-t)u_1(k) + t\underline{u}|}.$$

We refer to [4] for more details. An example of interpolation of a path is given on Figure 3.

We compare our algorithm in Figure 4 on a real material, silicon, to the standard method of Marzari and Vanderbilt, based on using projections as an initial guess for a minimization procedure. Our algorithm proves to be robust and converge to the same minimizer as the Marzari-Vanderbilt procedure with an appropriate initial guess, but without user input.

IV.7. Metallic systems: existence of localized Wannier functions

In the sections up to now we have developed the theory of Wannier functions assuming that we were interested in representing the spectral subspace associated with an isolated

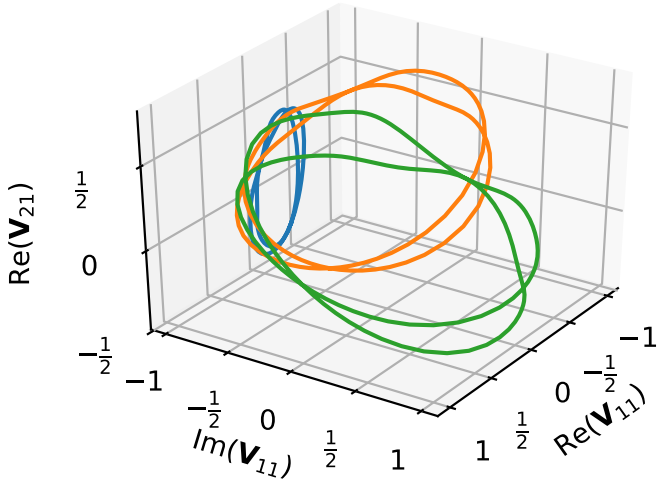


FIGURE 3. Deformation of a path in the sphere, used to construct Wannier functions for the Kane-Mele model (see [4]). The initial path ($t = 0$, green) is continuously deformed to a point ($t = 1/3$, yellow then $t = 2/3$, blue).

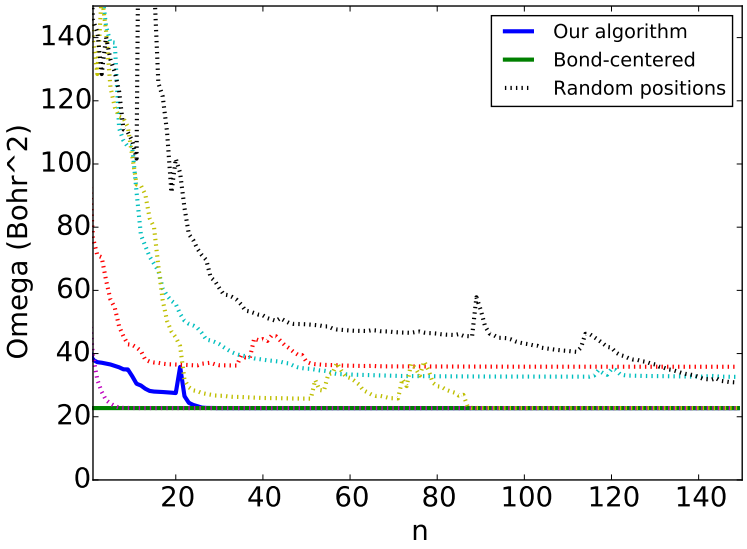


FIGURE 4. Total variance Ω (see (46)) as a function of the iterations in an iterative solver, in silicon, with different initial guesses: our algorithm (blue), bond-centered orbitals (green) and different random initial positions (different dashed lines).

set of eigenvalues $I \subset \mathbb{N}$:

$$\inf_{k\in\mathcal{B},i\in I,a\notin I} |\varepsilon_{ik} - \varepsilon_{ak}| > 0$$

This is the case for the occupied bands of insulators, and is the setting for the original definition of Wannier functions. However, it is often desirable to represent non-isolated sets of bands. As just one motivating example, the frequency-dependent conductivity tensor can be computed by the Kubo formula, yielding formulas such as [YWVS07]

$$\sigma_{\alpha\beta}(\omega) \propto \lim_{\delta\to 0^+} \int_{k\in\mathcal{B}} \sum_{n=1}^{N_k} \sum_{m=N_k+1}^{\infty} \frac{\langle u_{nk}|(-i\nabla_\alpha)|u_{mk}\rangle\langle u_{mk}|(-i\nabla_\beta)|u_{nk}\rangle}{(\varepsilon_{nk}-\varepsilon_{mk})(\varepsilon_{nk}-\varepsilon_{mk}-\omega+i\delta)} dk \tag{50}$$

where $N_k = |\{\varepsilon_{nk} \leq \varepsilon_F, n \in \mathbb{N}\}|$ with ε_F the Fermi level. This is a parametric integral of a discontinuous quantity that is expensive to evaluate. On the other hand, the low-frequency behavior is mainly dominated by the bands close to the Fermi surface, and it is therefore desirable to obtain an interpolation of these bands, which are generally not isolated from the rest of the spectrum.

Since the subspace $X(k) = \mathrm{Span}(u_{nk})_{n\in I}$ is not smooth, interpolating it directly is inefficient. However, if we can find a space $Y(k)$ of dimension $|I| + N_{\mathrm{extra}}$, with $N_{\mathrm{extra}} > 0$, such that $X(k)$ is a subspace of $Y(k)$, together with a smooth and periodic basis of $Y(k)$, then we can reconstruct the desired spectral information (such as eigenvalues $(\varepsilon_{nk})_{n\in I}$) of X. This is the basis for the "disentanglement" scheme proposed in 2001 [SMV01], which has found wide applicability in the computation of various properties [MMYSV12]. There, one optimizes the spread Ω of an extended set of functions $(v_{nk})_{n=1,\dots,|I|+N_{\mathrm{extra}}}$, subject to the constraint that the v_{nk} span the "frozen" subspace $X(k)$:

$$\mathrm{Span}(u_{nk})_{n\in I} \subset \mathrm{Span}(v_{nk})_{n=1,\dots,|I|+N_{\mathrm{extra}}}.$$

This will ensure that for instance the reduced Hamiltonian $\widetilde{H}_{mn}(k) = \langle v_{mk}, Hv_{mk}\rangle$ contains the $(\varepsilon_{nk})_{n\in I}$ as eigenvalues, and that Wannier interpolation of the ε_{nk} is possible.

In [6], we studied this problem theoretically. We show that, under generic assumptions satisfied in most cases, there exists such a set v_{nk} which can be chosen infinitely differentiable, and that $N_{\mathrm{extra}} = 1$ is enough.

THEOREM IV.7.1 (Existence of localized metallic Wannier functions [6]). *Assume that $d = 3$ and that the system has the time-reversal property $H_{-k} = CH_kC$, where C is an anti-unitary complex operator such that $C^2 = 1$. Let $I = \{1, \dots, N\}$, and*

$$K_N = \{k \in \mathcal{B}, \varepsilon_{Nk} = \varepsilon_{N+1,k}\}$$

the crossing set. Assume that K_N and K_{N+1} are finite unions of isolated points and piecewise smooth curves, and that $K_N \cap K_{N+1} = \emptyset$. Then there exists a smooth and periodic orthogonal family $(v_{nk})_{n=1,\dots,N+1}$ such that $\mathrm{Span}(u_{nk})_{n\in I} \subset \mathrm{Span}(v_{nk})_{n=1,\dots,|I|+1}$.

The family v_{nk} can be infinitely differentiable; however, it cannot be analytic in general.

We now sketch the arguments of the proof, in the simpler case where K_N and K_{N+1} are composed of isolated points. First, we build the projector $P(k)$ on the subspace $Y(k)$; second, we find a smooth and periodic basis v_{nk} of $\mathrm{Ran}\, P(k)$. In order to build P, a natural candidate is to start from $P_{N+1}(k) = \sum_{n=1}^{N+1} |u_{nk}\rangle\langle u_{nk}|$. This P_{N+1} spans the frozen subspace $\mathrm{Ran}\, P_N$, but it is discontinuous at $k \in K_{N+1}$, where $\varepsilon_{N+1,k}$ crosses with $\varepsilon_{N+2,k}$. It is therefore natural to try to take $P = P_{N+1}$ everywhere except in neighborhoods Ω_i of the points of K_{N+1}, and to continue it inside smoothly, as illustrated in Figure 5.

There is however a topological obstruction to this program, as can be seen in the following theorem, which extends Theorem IV.4.1

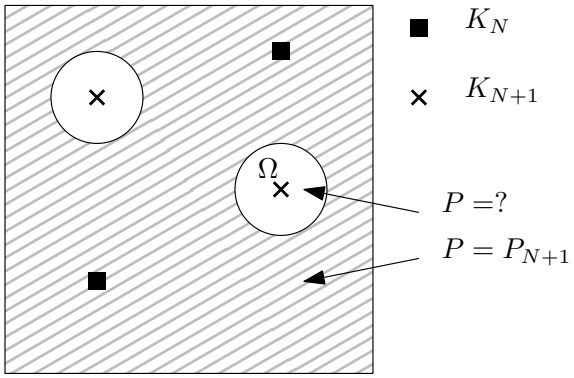


FIGURE 5. A two-dimensional visualization of the Brillouin zone, with $P = P_{N+1}$ everywhere except in a neighborhood of K_{N+1}

THEOREM IV.7.2. *Let Ω be an open set in $\mathbb{R}^3$, $S = \partial\Omega$ be a smooth, compact, oriented two-dimensional surface, and $S \ni k \mapsto X(k)$ be a smooth complex finite-dimensional subspace on S. Then the following are equivalent:*

- *There exists a continuous orthogonal basis $v(k)$ of $X(k)$ on S;*
- $\mathrm{Ch}(S, X) = 0$*;*
- *There exists a continuous extension of X inside Ω, i.e. a map $\overline{\Omega} \ni k \mapsto \widetilde{X}(x)$ such that $\widetilde{X}(k) = X(k)$ on S.*

Again, this theorem is classical; see [6] for a hands-on proof.

Coming back to our favorite example

$$H(k) = k \cdot \sigma = \begin{pmatrix} k_3 & k_1 - ik_2 \\ k_1 + ik_2 & -k_3 \end{pmatrix} \tag{51}$$

on the sphere $S = \{k \in \mathbb{R}^3, |k| = 1\}$, if $X(k)$ is the subspace associated to the lowest eigenvalue -1 of $H(k)$, we have $\mathrm{Ch}(S, X) = 1$. It is easy to understand why there cannot exist a continuous extension $\widetilde{X}$ of X: if there was, one could take an arbitrary basis $v(0)$ of $\widetilde{X}(0)$, and extend it using parallel transport along the radial direction to a basis of $X(k)$ on S, contradicting Theorem IV.4.1.

Since K_{N+1} generically contains conical intersections of the form (51) above, $\mathrm{Ch}(P_{N+1}, \Omega_i) \neq 0$ on each individual neighborhoods Ω_i of K_{N+1}. This means that P_{N+1} cannot be continued inside the individual neighborhoods Ω_i as in Figure 5. Rather we exploit the fact that Chern numbers are additive, and that the sum of the Chern numbers associated with the crossings in K_{N+1} must vanish. Recall that

$$\mathrm{Ch}(S, X) := \frac{1}{2\pi} \int_S \mathcal{F}_X, \tag{52}$$

where the 2-form

$$\mathcal{F}_X = -i \operatorname{Tr}(P_X dP_X \wedge dP_X)$$

is the Berry curvature. As a curvature, this form satisfies $d\mathcal{F}_X = 0$ (when seen as a vector field in three dimensions, this means that its divergence is zero). Let Ω be a connected

set that covers all the K_{N+1}, while excluding the K_N (see Figure 6 for a simple case; the general construction is detailed in [6]). Then, using Stokes' theorem, we have

$$\int_{\partial\Omega} \mathcal{F}_X = \int_{\mathcal{B}\setminus\Omega} d\mathcal{F}_X = 0.$$

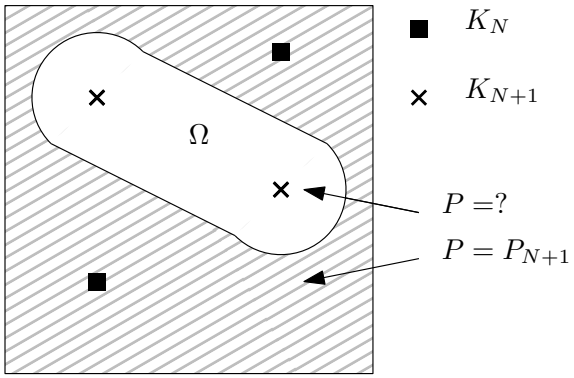


FIGURE 6. The Brillouin zone, and the domain Ω on which we perform the extension.

It follows that P_{N+1} can be continued inside Ω. It remains to make sure that this continuation still contains the frozen subspace X_N; this is done by continuing $p = P_{N+1} - P_N$ inside Ω while maintaining the orthogonality of $\operatorname{Ran} p$ and $\operatorname{Ran} P_N$, and reconstructing P as $P_N + p$. Finally, a basis v_{nk} of $P(k)$ must be constructed, which requires the extension to be performed in a fashion that is compatible with time-reversal. We refer to [6] for details.

IV.8. Metallic systems: the Marzari-Vanderbilt procedure

As mentioned in Section IV.2, the standard procedure to build localized Wannier functions, both for isolated and non-isolated bands, is to minimize the spread functional

$$\Omega = \sum_{n=1}^{N_{\text{wan}}} \left(\int_{\mathbb{R}^3} |x|^2 |w_{n0}|^2(x) dx - \left| \int_{\mathbb{R}^3} x |w_{n0}|^2(x) dx \right|^2 \right), \tag{53}$$

an implicit functional of the $(v_{nk})_{n=1,\dots,N_{\text{wan}}}$.

In the case of isolated bands, $N_{\text{wan}} = |I|$, and

$$v_{nk} = \sum_{m \in I} u_{mk} U_{mn}(k)$$

for some family of unitary matrices $U(k)$. The minimization of Ω is done in practice by a Riemannian conjugate gradient algorithm [AMS09] similar to the methods described in Section II.7.1. This proves to be robust, as long as a suitable initial guess is chosen. The computational efficiency of these methods is of limited importance, as their cost is dwarfed by that of the computation of the u_{nk}.

The construction for non-isolated band structures is more complex. There, $N_{\text{wan}} > |I|$:

$$v_{nk} = \sum_{m \in \mathbb{N}} u_{mk} U_{mn}(k)$$

where $(U_{mn}(k))_{m\in\mathbb{N}, n\in I}$ is a semi-infinite matrix (of course, in practice, the summation on m is truncated) with orthogonal columns. The constraint that the bands $(\varepsilon_{nk})_{n\in I}$ are to be reproduced exactly is

$$\operatorname{Span}(u_{nk})_{n\in I} \subset \operatorname{Span}(v_{nk})_{n=1,\dots,|I|+N_{\text{extra}}}.$$

To minimize the spread Ω, the procedure in [SMV01], routinely used in practice, is to split the functional $\Omega = \Omega_I + \Omega_D$ into a gauge-independent part Ω_I, that only depends on the subspace $P(k)$ (equivalently, on $U(k)U(k)^*$), and a gauge dependent part Ω_D, that depends on the choice of basis v_{nk}. The next step is to optimize Ω_I with respect to $P(k)$, fix this $P(k)$, then optimize the basis v_{nk} of $\operatorname{Ran} P(k)$. The advantage of this procedure is that it splits into two familiar problems. The first step is to optimize a functional under the constraint that $P(k)$ is a projector. This is done by reformulating the opotimality conditions as a nonlinear eigenvector problem similar to the one of Kohn-Sham density functional theory. The second is a minimization under orthogonality constraint of the same form as in the insulating case.

Note that there is no reason to expect that the minimizer obtained with this two-step procedure is a minimum of the original functional Ω. In [**5**], we proposed to use a different parametrization of U as the product of two matrices with orthogonal columns to fully optimize Ω. The results obtained differ slightly from the two-step procedure, but retain the same qualitative character.

This full minimization procedure allows us to consider numerically the question: what are the localization properties of the metallic maximally-localized Wannier functions, i.e. the minimizers of the problem

$$\begin{aligned} \min \quad & \Omega \\ \text{s.t.} \quad & \langle v_{mk}, v_{nk} \rangle = \delta_{mn}, \\ & \operatorname{Span}(u_{nk})_{n\in I} \subset \operatorname{Span}(v_{nk})_{n=1,\dots,|I|+N_{\text{extra}}}. \end{aligned}$$

In the case of insulators, this question was studied in [PP13], where it is proven that the minimizers are exponentially localized. Heuristically, this is because, since multiplication by x in real space is equivalent to differentiation in reciprocal space, Ω acts as a kind of Dirichlet energy. By analytic regularity, the minimizers are then analytic in reciprocal space, and therefore exponentially localized in real space.

In [**5**], we apply this procedure to the case of the free electron gas, i.e. the operator $-\Delta$ seen as a periodic operator with period 2π. The spectrum in this case consists in bands $|k+K|^2$ for all $K \in \mathbb{Z}$, $k \in [0,1]$. At $k = 0.5$, the first band $\varepsilon_{1k} = \min(|k-K|^2, K \in \mathbb{Z})$ intersects with the second. We seek to build two Wannier functions that reproduce exactly the first band. For this operator, this means that we seek two localized functions w_1 and w_2 such that $(w_i(\cdot - R))_{i\in\{1,2\},\, R\in 2\pi\mathbb{Z}}$ is an orthogonal basis, that spans all L^2 functions whose Fourier transform is supported in $|\xi| \leq \frac{1}{2}$. Note that the cardinal sine $\sin(x)/x$ provides such a basis, although very weakly localized (non-integrable).

The minimizers of the spread functional Ω can be visualized Figure 7.

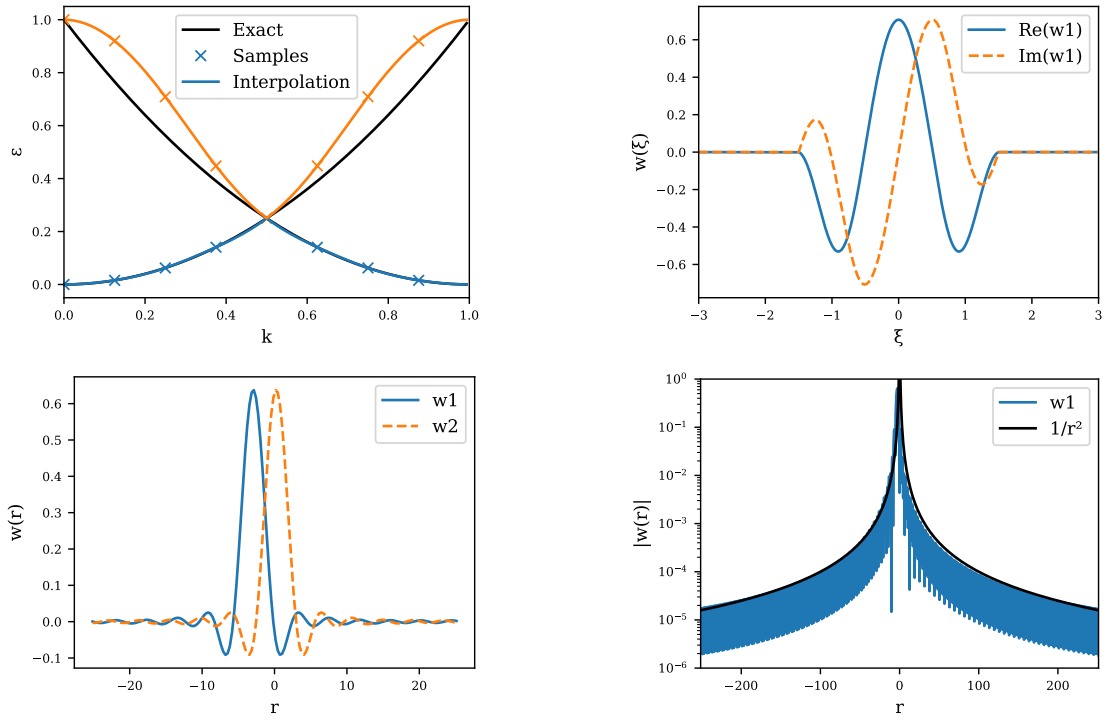


FIGURE 7. Wannier interpolation of the one-dimensional free electron gas (top left) with Wannier functions in reciprocal space (top right). Wannier functions in real space, in linear (bottom left) and logarithmic (bottom right) scale.

The Wannier functions are smooth everywhere in reciprocal space, except at $\xi = \pm\frac{3}{2}$ where they have cusps. This is because the constraint $\mathrm{Span}(u_{1k}) \subset \mathrm{Span}(v_{1k}, v_{2k})$ is discontinuous at $k = \frac{1}{2}$, so v_k cannot be expected to be differentiable at $k = \frac{1}{2}$. In real space, this implies an oscillatory tail, decaying as $1/r^2$. In this particular 1D example, it is easy to fix this decay by smoothing the v_{nk} in reciprocal space. We then obtain Wannier functions that decay faster than any inverse polynomial (see [5]). This construction can be generalized to higher dimensions, yielding localized shift-orthogonal bases of band-limited functions, a sinc-like construction reminiscent of wavelets.

This example can be generalized to higher dimensions. It shows that, in general, the minimization of the Marzari-Vanderbilt spread for entangled band structure does not yield Wannier functions that are more than algebraically localized, although we know from the theory in Section IV.7 that there exist Wannier functions decaying faster than any inverse polynomial.

IV.9. Perspectives

In theory, our construction for insulators perfectly answers the question of how to construct smooth bases. In practice, it is found to work very well for crystals with a small

unit cell and dense k-point mesh, but to yield no useful information for crystals with large unit cells. This is because we only try to localize the Wannier functions across unit cells, not inside (remember that we started with an arbitrary unitary matrix at $k = 0$). In practice, the method of [DLY15], based on picking the set of initial projections as delta functions in a greedy manner, is found to perform very well across a wide range of materials [VPMYM+19], but fails on complicated band topology such as the Kane-Mele model. It would be useful to find a method that has the advantages of both our method and that of [DLY15].

Regarding metallic systems, the localization properties of minimizers of the Marzari-Vanderbilt Ω criterion on realistic systems are still unclear. It would be interesting to explore whether they can be improved by a modified scheme. Finally, high-order eigenvalue interpolation beyond the framework of Wannier functions with a rigidly frozen window is a topic worthy of further inquiries.

CHAPTER V

Iterative methods for molecular simulation

In this chapter, I present collaborative works that focused on improving iterative methods, with application areas ranging from condensed matter physics to biochemistry and theoretical physics. The four sections are independent and correspond to four different papers: [12], [8], [10], [9]. This chapter is less unified than the others, and is more computational in flavour.

V.1. Linear eigenvalue problems: plane-wave DFT

In [12], we worked on the massive parallelization of the eigensolver in the ABINIT code, which solves the equations of KSDFT in a plane wave basis. In the SCF framework (see Section II.7.2), the inner loop consists in solving the equation

$$\left(-\frac{1}{2}(-i\nabla + k)^2 + V\right) u_{nk} = \varepsilon_{nk} u_{nk}, \tag{54}$$

with $\langle u_{nk}, u_{nk}\rangle = \delta_{mn}$ in a unit cell (we will take $[0, 2\pi]^3$ without loss of generality), with periodic boundary conditions. The self-consistent potential V also includes non-local terms originating from pseudopotentials (see Section II.6). When expanded in a plane-wave basis, we are left with an eigenvalue problem

$$Ax_i = \lambda_i x_i,$$

where A is a matrix of size $N_{\text{pw}} \times N_{\text{pw}}$.

This system is then to be solved for the first N_{el} electrons. In practice, in a plane-wave basis, the basis size N_{pw} is larger than N_{el} by a factor of $100 - 1000$. For large systems, it is infeasible to store the matrix A, and iterative methods must be used. This is greatly facilitated by the special structure of A, which can be applied to a vector efficiently using Fast Fourier transforms. Therefore, although A is not sparse, the cost of applying A to a vector is $O(N_{\text{pw}} \log N_{\text{pw}})$ (at least when non-local pseudopotentials are not used).

Iterative eigensolvers have a long history, and a variety of algorithms exist [Saa11]. For our purposes, we need an eigensolver that

- Operates in a matrix-free fashion (only requires matrix-vector products with A);
- Can find the first N_{el} eigenvectors of a N_{pw} matrix, with $N_{\text{el}} \ll N_{\text{pw}}$;
- Can use a preconditioner (A is strongly diagonally dominant for large frequencies).

One solver that fits into this category is the Locally Optimal Block Preconditionned Conjugate Gradient (LOBPCG) algorithm. This algorithm, like most iterative eigensolvers, is based on the Rayleigh-Ritz procedure: given a set of vectors $X = (x_i)_{i=1,\dots,N_{\text{RR}}}$, find $Y = (y_i)_{i=1,\dots,N_{\text{el}}}$ with orthogonal columns such that $\text{Span}Y \subset \text{Span}X$, and $\text{Tr}(Y^* A Y)$ is minimum. If the columns of X are linearly independent, this can be solved by computing the $N_{\text{RR}} \times N_{\text{RR}}$ matrices

$$A_X = X^* A X, \qquad O_X = X^* X,$$

solving the generalized eigenvalue problem

$$A_X c = \lambda O_X c \tag{55}$$

for the N_{el} lowest eigenvalues, and returning $Y = Xc$. We call this procedure $Y = \texttt{RR}(X)$.

In its basic version, without preconditioning and in exact arithmetic, the LOBPCG algorithm can be written as

$$X_{n+1} = \texttt{RR}(X_n, AX_n, X_{n-1}).$$

There are numerous implementation details that have to be taken care of in the implementation of this algorithm: preconditioning, method of orthogonalization, locking of converged vectors... In particular, numerical stability is very challenging to ensure [DSYG18; HL06].

In a parallel setting, the advantage of this algorithm is that the products Ax_i can be done in parallel for all $i = 1, \ldots, N_{\text{el}}$. This is important because spatial parallelization of the matrix-vector product is limited (FFTs have poor parallel scalability). However, the Rayleigh-Ritz procedure involves solving (55). This requires communication between different eigenvectors, which slows down performance for a large number of processors.

In [**12**], we implemented Chebyshev filtering in the ABINIT code and studied its parallel scalability. Chebyshev filtering, introduced for KSDFT in [ZSTC06] based on work done in the '70s [Rut70], is based on the iteration

$$X_{n+1} = \texttt{RR}(p(A)X_n),$$

where p is a polynomial designed to be large on the wanted eigenvalues, and small on the unwanted ones, as shown in Figure 1. The polynomial $p(A)$ of degree N_p can be applied to the vectors in X_n by N_p applications of the matrix A. Increasing the degree yields a better separation between the wanted and unwanted subspaces, at the cost of more applications of the matrix A.

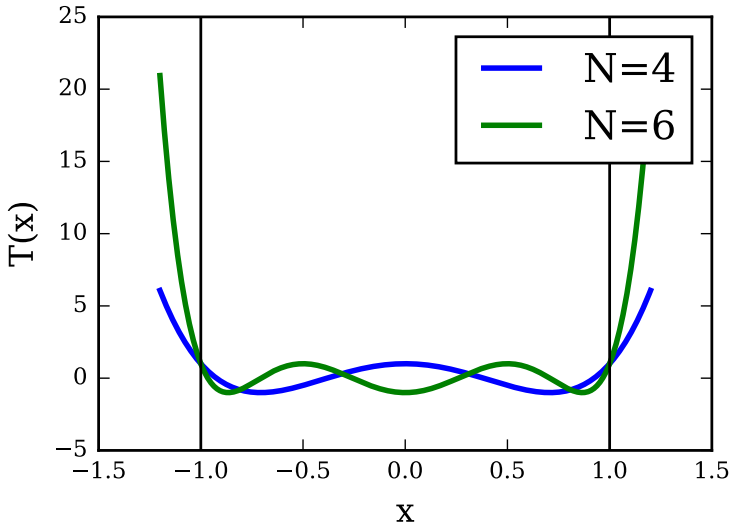


FIGURE 1. Chebyshev polynomials of degrees 4 and 6, small on $[-1, 1]$ and large outside. By shifting this polynomial it is possible to ensure that $T(x)$ is large on $[\lambda_1, \lambda_{N_{\text{el}}}]$ and small on $[\lambda_{N_{\text{el}+1}}, \lambda_{N_{\text{pw}}}]$.

The advantage of this algorithm compared to LOBPCG is that it requires less Rayleigh-Ritz steps to converge, and that the bases involved are smaller (N_{el} against $3N_{\text{el}}$). The trade-off is a higher number of applications of A and the inability to use preconditioning, which was found to be acceptable in some scenarios in our tests. Since the Chebyshev method uses less communication between eigensolvers, its parallel scalability is better. We tested this on the Curie supercomputer of TGCC, demonstrating scalability to tens of thousands of processors (see Figure 2).

A complication appears in the treatment of advanced methods for pseudopotentials such as the Projector Augmented-Wave (PAW) method [Blö94]. There, a generalized eigenvalue problem $Ax = \lambda Bx$ has to be solved, with B a Hermitian positive definite matrix. To apply the Chebyshev method, one has to form polynomials in $B^{-1}A$, which can be costly to perform. In [**12**], we used the fact that B is a low-rank perturbation of the identity together with the Woodbury matrix identity to reduce this inversion to solving a smaller-size linear system. We then solved this system with a preconditioned iterative method.

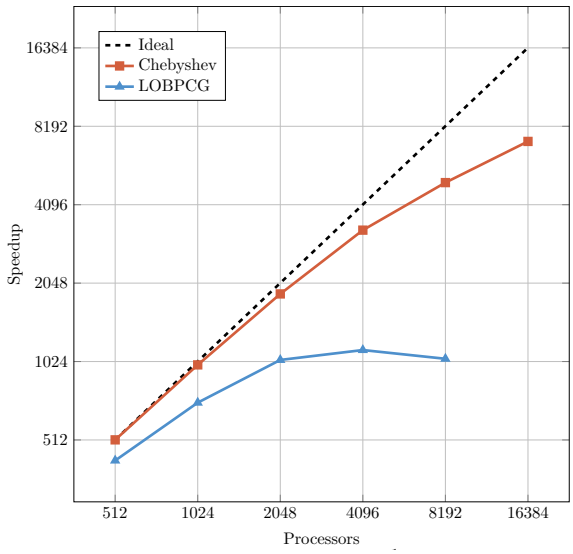


FIGURE 2. Parallel scalability $\left(\frac{T_{\text{serial}}}{T_{\text{parallel}}}\right)$ of the Chebyshev filtering algorithm against LOBPCG.

V.2. Nonlinear eigenvalue problems: Bose-Einstein condensation

In [**8**], we worked on minimization algorithms for the efficient computation of Bose-Einstein condensates.

Bose-Einstein condensation is a peculiar state of matter that happens when a dilute gas of bosons (for instance, atoms with an even mass number) is cooled below a certain temperature. In this regime, the bosons all occupy the same quantum state. They are of interest in fundamental physics because they are macroscopic objects with quantum properties.

The simplest ansatz for the wave function of such a system of N bosons is the wave function

$$\psi(x_1, \dots, x_N) = \phi(x_1) \dots \phi(x_N)$$

where $\phi \in L^2(\mathbb{R}^3, \mathbb{C})$ is a normalized one-particle wave function. Assuming only contact interactions between the particles and plugging this ansatz into the Schrödinger equation yields the Gross-Pitaevskii equation [PS08]:

$$E(\phi) = \int_{\mathbb{R}^3} \frac{1}{2}|\nabla\phi|^2 + V|\phi|^2 + \alpha|\phi|^4,$$

where V is the external potential, and α is the interaction strength.

The associated minimization problem

$$\inf_{\|\phi\|=1} E(\phi) \tag{56}$$

has Euler-Lagrange equations

$$\left(-\frac{1}{2}\Delta + V + 2\alpha|\phi|^2\right)\phi = \lambda\phi, \quad \|\phi\| = 1.$$

The Gross-Pitaevskii equation is mathematically interesting because it is a simplified version of the problem of KSDFT, with only one orbital. As such, it has been used to prototype mathematical approaches for electronic structure [CDMSV14]. It has however different mathematical properties: for instance, when the equation is coupled with a magnetic field the Aufbau principle does not usually hold (λ is not the smallest eigenvalue of $-\frac{1}{2}\Delta + V + 2\alpha|\phi|^2$), and the methods used to solve it are different.

In numerical simulations of Bose-Einstein condensates, the potential V is often taken to be a harmonic trap $V(x) = \frac{1}{2}|x|^2$, to simulate confinement. The simulation is truncated to a finite box of size L, then ϕ is discretized using finite differences, finite elements or a spectral (plane-wave) method; in [**8**], we focused on the latter. In all setups, we have to solve a discrete problem of the form (56), with $\phi \in \mathbb{R}^{N_b}$. Previous methods for this minimization problem focused on methods inspired by the imaginary time method: the time-dependent Schrödinger equation in imaginary time becomes a gradient flow for (56). This can then be discretized using methods usual in evolution partial differential equations, such as backward Euler of the Crank-Nicolson scheme [BC12]. However, these indirect methods are inefficient when applied to a minimization problem: one has to solve a linear system at each step, incurring the cost of a Newton scheme without its convergence rate. In [**8**], we instead used a simple nonlinear conjugate gradient method to solve the minimization problem. The normalization constraint is taken into account by methods of Riemannian optimization [EAS98].

A major ingredient in a minimization method is preconditioning. A good preconditioner for a PDE should take into account the nature of the continuous problem [MS14]. In particular consider the problem of preconditioning the operator

$$-\frac{1}{2}\Delta + \frac{1}{2}|x|^2$$

on $\mathbb{R}^3$, which is the dominant part of the Hessian of our problem. There are two sources of divergence here: first, the unboundedness of the Laplacian on high-frequency modes; and second, the unboundedness of $|x|^2$ on spatially extended modes (these two problems being

dual under the Fourier transform). This manifests in an unpreconditionned minimization algorithm by an increased number of iterations as the discretization parameters are refined: if L is the size of the box and h the grid size, the spectral radius of the matrix is $O(L^2+h^{-2})$.

Note that both $-\Delta$ and $|x|^2$ are trivial to precondition separately, since these operators are diagonal either in reciprocal or direct space. If preconditionning with $-\Delta$ (also called "Sobolev gradient method" [Neu09], since it is equivalent to taking the gradient of E with respect to the Sobolev H^1 metric), the conditionning is $O(L^2)$; if preconditionning with V, the conditionning is $O(h^{-2})$. In **[8]**, we conducted extensive numerical experiments confirming this behavior, and experimented with combined preconditionners. Finding an efficient preconditionner that is able to make this conditionning mesh-independent is an important open problem.

V.3. Linear systems: polarizable force fields

In **[10]**, we worked on the implementation of polarizable force fields to accelerate the simulation of large biochemical systems.

A force field is a function $F : \mathbb{R}^{3N} \to \mathbb{R}$ mapping atomic configurations to their total potential energy, in the Born-Oppenheimer approximation. For a given atomic configuration x, $F(x)$ can be computed from the solution of an electronic structure problem (see Chapter II), or can be approximated through the use of force fields. The simplest such force field is the Lennard-Jones potential

$$F(x) = \sum_{1 \le i < j \le N} V_{\mathrm{LJ}}(|x_i - x_j|)$$

with

$$V_{\mathrm{LJ}}(r) = 4\varepsilon \left(\left(\frac{r_0}{r} \right)^{12} - \left(\frac{r_0}{r} \right)^{6} \right)$$

for some characteristic energy ε and length r_0. While this potential gives accurate results for systems dominated by dispersion effects (typically, noble gases), it fails to reproduce the behavior of more complex systems, in particular those in which covalent bonding and electrostatics is important. For this, more accurate force fields are needed.

In classical force fields such as CHARMM [BBIMJNP+09] and AMBER [CCIDGL+05], developed in the '80s and used continuously since, covalent bonding is usually treated by harmonic terms. The total energy F contains terms of the form $\frac{1}{2}k(l - l_0)^2$, where l is the distance between two covalently-bonded atoms, l_0 is their equilibrium distance and k is a force constant. A similar energy term is also added on bond and dihedral angles. The electrostatics is usually modeled by partial charges: for instance, a water molecule might be represented by a negative charge $-q$ on the oxygen atom, and a positive charge $q/2$ on each of the hydrogen atoms. The parameters in the model $(k, l_0, q, \dots)$ are fitted to experimental or quantum-mechanical data. This type of classical force fields has been shown to be accurate for the prediction of many properties, and is now a workhorse of computational chemistry and biochemistry.

These classical force fields, however, neglect a number of important effects such as induced polarization: when subjected to an applied electric field E, the charge distribution in an atom changes (the atom polarizes). Force field models taking this effect into account,

called polarizable force fields, have been developed and improve accuracy significantly [PWRPC+10].

The simplest model for the change in the distribution of charge of an atom is an induced dipole $\mu = \alpha E_{\text{tot}}$, where α is the polarizability, a 3×3 positive definite matrix, and E_{tot} is the total electric field. The electric field E_{tot} is itself affected by the induced dipoles of other atoms, the long-range field created by a single dipole μ located at the origin being

$$E_\mu(x) = \frac{3(\mu \cdot x)x - |x|^2\mu}{4\pi\varepsilon_0|x|^5},$$

with ε_0 the permitivity of the vacuum. For the atom i in a collection of N atoms with positions $(x_i)_{i=1,\dots,N}$ in an external field E, we have therefore

$$\mu_i = \alpha_i E_{\text{tot}} = \alpha_i \left(E(x_i) + \sum_{j \neq i} E_{\mu_j}(x_i - x_j) \right) = \alpha_i E(x_i) + \sum_{j \neq i} \alpha_i T_{ij} \mu_j,$$

where the tensor T_{ij} is defined through $E_{\mu_j}(x_i - x_j) = T_{ij}\mu_j$. This can then be reformulated in vector form as

$$T\mu = E \tag{57}$$

where $E_i = E(x_i)$, and

$$T = \begin{pmatrix} \alpha_1^{-1} & -T_{12} & -T_{12} \dots & T_{1N} \\ -T_{21} & \alpha_2^{-1} & -T_{23} \dots & T_{2N} \\ -T_{31} & -T_{32} & \ddots & \\ \vdots & \vdots & & \vdots \\ -T_{N1} & -T_{N2} & \dots & \alpha_N^{-1} \end{pmatrix}.$$

This equation has a major flaw: the matrix T is not necessarily positive definite, and therefore the associated energy is not bounded from below, a phenomenon known as the "polarization catastrophe". This is because the dipole approximation is only valid at large distances, and produces non-physical results when extended to smaller distances. This is usually fixed in an ad-hoc way by modifying the form of T_{ij} at small distances to ensure positive-definiteness ("Thole damping") [Tho81].

To compute the energy due to polarization effects associated to a given configuration (which is added to the bonded, dispersion and electrostatic energies), the procedure is to solve (57) for the dipoles μ in the electric field E generated by the permanent charges in the system, and then compute the energy associated to these dipoles:

$$\mathcal{E}(\mu) = \frac{1}{2}\mu^T T \mu - E^T \mu. \tag{58}$$

The force $\frac{d\mathcal{E}}{dx}$ can then be computed from the Hellmann-Feynman type expression

$$\frac{d\mathcal{E}}{dx} = \frac{\partial \mathcal{E}}{\partial \mu} \cdot \frac{\partial \mu}{\partial x} + \frac{\partial \mathcal{E}}{\partial x} = \frac{\partial \mathcal{E}}{\partial x} = \frac{1}{2}\mu^T \frac{dT}{dx}\mu - \mu^T \frac{dE}{dx} \tag{59}$$

where we have used (57) to show that $\frac{\partial \mathcal{E}}{\partial \mu}$ is zero. In practice, (57) is solved using the conjugate gradient method, the matrix T being very well conditioned. When the system

is simulated with periodic boundary conditions, as is often the case, the application of T to a vector can be performed using the efficient smooth particle-mesh Ewald (SPME) method, which uses fast Fourier transforms to compute the long-range summations.

In the large molecular systems found in biology, the solution of (57) is the most time-consuming step of the computation. It is therefore crucial to keep the number of iterations of the iterative solver to a minimum. However, when (57) is solved only approximately, the formula (59) for the force is also approximate. When performing an NVE simulation (integrating Newton's laws of motion for the potential energy $\mathcal{E}$), this results in a force which is not conservative, and therefore an energy drift for large times. To avoid this, [WS05] proposed to solve approximately (57) by a fixed number of iterations of the block Jacobi method. This results effectively in the approximation

$$\mu_{\mathrm{WS}} = p(\alpha T)E$$

where $\alpha = \mathrm{diag}(\alpha_1, \ldots, \alpha_N)$ and p is a fixed polynomial. From that expression $\mu_{\mathrm{WS}}(E)$, one can compute the exact forces corresponding to the energy $\mathcal{E}(\mu_{\mathrm{WS}}(E))$ by computing explicitly the gradient of (58), without using the Hellmann-Feynman expression. In [SPISCIB15], the authors improved this method by interpolating empirically the results of [WS05] between different orders.

In **[10]**, we proposed to instead solve (57) by a fixed number of iterations of a conjugate gradient method. Although the expressions of the energies and forces are more complex, they can still be handled efficiently, and the resulting method gives accurate results while being free of empirical parameters. In retrospect, our method can be seen as a manual implementation of the methodology of reverse-mode automatic differentiation [Gri+89].

V.4. Saddle point search: reaction paths

In **[9]**, we investigated the convergence of algorithms for saddle point search. Such algorithms are important for instance to compute quantities dictating the kinetics of chemical reactions in the gas and liquid phases, and of atom migration in solids.

Let $F : \mathbb{R}^N \to \mathbb{R}$ be the potential energy surface (obtained through quantum-mechanical calculations or by a classical force field) of a molecular system. By the principles of statistical physics, each state $x \in \mathbb{R}^N$ has a probability density proportional to $e^{-\frac{F(x)}{k_B T}}$. It follows that, at low temperature, the states with the largest occupation are the energy minima. These correspond to equilibrium states of the system. Chemical reactions correspond to transitions between these states, wherein thermal fluctuations move the system between different local minima. Chemical kinetics is the study of the rates at which this process occurs. The simplest theory able to predict these rates is transition state theory, as illustrated in Figure 3.

In this theory, the transition rate between two states x_A and x_B is given by the formula

$$k = Ae^{-\beta E}$$

where A is a prefactor, and

$$E = \min_{\gamma:[0,1]\to\mathbb{R}^N, \gamma(0)=x_A, \gamma(1)=x_B} \max_{t\in[0,1]} F(\gamma(t)) \tag{60}$$

is the barrier height. This minimum is achieved at a saddle point x_{AB}, which is called the transition state.

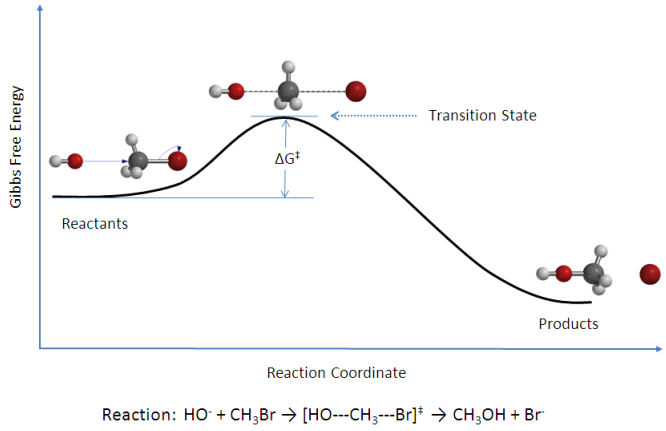


FIGURE 3. Transition state theory of a chemical reaction. Source: https://en.wikipedia.org/wiki/Transition_state_theory.

To predict chemical reactions and their reaction rates, one must therefore compute saddle points. From the mountain pass lemma [AE06], the variational problem (60) is mathematically well-posed given two local minima x_A and x_B. Accordingly, several algorithms such as the Nudged Elastic Band (NEB) or the string method attempt to discretize the reaction path γ directly [JMJ98; ERVE02]. This however only applies when the starting and end points are known. It is often of interest to explore a potential energy surface starting from only a single state x_A, i.e. escape from the potential well of x_A. For this, the method of choice is often the dimer method.

An idealized version of this method can be formulated as the ordinary differential equation

$$\dot{x} = -(1 - 2v_1(x)v_1(x)^T)\nabla F(x), \tag{61}$$

where $v_1(x)$ is the eigenvector associated with the lowest eigenvalue of $\nabla^2 F(x)$. This method acts as a gradient descent in the directions orthogonal to $v_1(x)$, and as a gradient ascent in the direction of $v_1(x)$. This makes it locally convergent near a simple saddle point x_{AB} where $\nabla^2 F(x_{AB})$ has a single negative eigenvalue. In particular, it is tempting to see (61) as the analogue of gradient flows for saddle points instead of minima, and expect good convergence properties for this flow to a saddle point. Work along these lines has been initiated in [GOP16], where the authors use a merit function to try to increase the robustness of the dimer method.

In practice, the equation (61) is discretized with a forward Euler method, and the eigenvector $v_1(x)$ is computed with a gradient descent on the Rayleigh quotient, where the matrix-vector product $\nabla^2 F(x)v$ is performed using finite differences, only requiring the user to provide a way to compute ∇F. We idealized this by the flow

$$\begin{cases} \dot{x} &= -(1 - 2vv^T)\nabla F(x) \\ \varepsilon \dot{v} &= -(1 - vv^T)\nabla^2 F(x)v \end{cases} \tag{62}$$

where the second equation represents the gradient descent on the Rayleigh quotient, and ε is a parameter controlling the speed at which v converges to $v_1(x)$. As $\varepsilon \to 0$, we recover formally (61).

In [**9**], we have investigated the differential equations (61) and (62) from a theoretical point of view. We proved estimates for their basin of convergence to a saddle point, but also highlighted several counter-examples to convergence. These are based on the study of singularities of the flow (61), which are points where $\nabla^2 F(x)$ has a degenerate lowest eigenvalue. In particular, we showed that, depending on the properties of $\nabla^3 F(x)$, these singularities can be attractive for the flow (61), in which case the solution ceases to exist in finite time. We showed that (62) then has limit cycles of size $\sqrt{\varepsilon}$ around these singularities. Using the von Neumann-Wigner theorem on the genericity of eigenvalue crossings [NW29], we showed that these singularities are stable under perturbations of F in any dimension (Theorem 8 in [**9**]).

Based on the concept of attractive singularities, we also designed as counterexamples potential energy wells from which both dynamics (61) and (62) can never escape, as shown in Figure 4. This example can be composed to yield potential energy surfaces with saddle points which are very hard to find with any method based on eigenvectors of the Hessian.

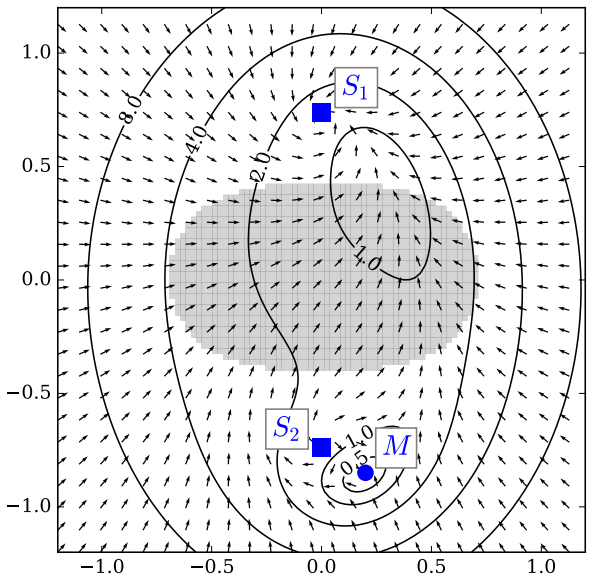


FIGURE 4. A potential well F from which (61) and (62) can never escape. The arrows represent the flow of (61). The point M is the minimum of F, S_1 and S_2 are attractive and repulsive singularities respectively.

The conclusion of that study is that while the dimer method can sometimes be used to explore potential energy surfaces, it can not be relied upon to converge globally to a saddle

point. Fundamentally, this is not surprising: while every coercive continuous function has a minimum, it does not necessarily have a saddle point. Therefore, in the absence of two local minima, from which one can infer the existence of a saddle point by the mountain pass lemma, it is vain to look for a saddle point that might not exist. The dimer method is then best used as a local method, combined with a globalization strategy, such as starting it from different initial positions.

CHAPTER VI

Perspectives

I present here some research tracks that I intend to explore in future work. Some of these correspond to work already initiated, others represent longer-term prospects.

VI.1. Ground-state computations

Ground-state computations with KSDFT is by now well-established. However, there are still numerically challenging systems, and the mathematical analysis of the algorithms is still very incomplete. Furthermore, as the field moves away from one-off computations towards high-throughput screening of materials or the generation of large databases, it is desirable to fully automatize numerical methods that currently require manual tuning.

VI.1.1. Minimization and self-consistent schemes. For a given discretization scheme, what is the optimal method for solving the Kohn-Sham equations? Although numerical practice has mostly crystallized around self-consistent iterations accelerated with the Anderson scheme (see II.7.2), there is reason to believe that this is suboptimal. First, this method lacks robustness, being guaranteed to converge to a local minimum of the energy only in setups that are not realized in practice (for instance, a large gap, or a very small damping parameter). Anderson acceleration can diverge unless very special care is taken in its implementation; this in turn degrades its efficiency. Furthermore, some systems are hard to converge, such as those with mixed insulating/metallic characteristics.

Direct minimization is a promising alternative, being more robust by construction. However, it is usually presented as being slower than self-consistent algorithms. With Eric Cancès and Gaspard Kemlin (PhD student at CERMICS/Inria, starting in October 2019), we are currently investigating theoretically the convergence rates of both types of algorithms, to be able to compare them on an equal footing (a paper is in preparation). I plan to then use this analysis to improve practical implementations of the algorithms.

Systems with degenerate Fermi levels, including metals, pose a distinct challenge to numerical methods. This problem is usually remedied by adding a small temperature to the model, but this is unsatisfactory as it changes the model and introduces an error. Instead, we should solve the system with fractional occupation numbers directly. There is previous work on this [CKST03; FBN09], but there are still issues both with the numerical analysis and the practical implementation. I plan to pursue this with Eric Cancès, Julien Toulouse (Sorbonne Université) and Zsuzsanna Tòth, who started as a postdoc in January 2020.

The convergence of the SCF method for large homogeneous systems of insulators and metals is now well-understood (see Chapter III Section III.3.4): the convergence rate is independent of the supercell size for insulators without preconditioning, and for metals with Kerker preconditioning. For hybrid systems containing both metallic and insulating parts, the convergence is found to be slow in practice. Despite some work in that direction (see e.g. [LY13]), a robust and completely automated procedure is still lacking. We are currently investigating this with Michael Herbst, who started as a postdoc in January 2019. Preliminary results, based on simple but flexible approximation of the dielectric operator, are very encouraging.

VI.1.2. Error analysis. The sources of error in standard plane-wave ground-state KSDFT computations are: the modeling error (frozen environment, neglected relativistic effects...), the DFT error (use of an approximate functional), the pseudopotential error, the discretization error (including both the plane-wave and k-points error), the truncation error (of nonlinear and linear solvers) and the floating-point arithmetic error. The modeling and DFT errors are extremely hard to control; by contrast the discretization, truncation and floating-point arithmetic errors are controlled approximations, in the sense that they can be systematically improved. The pseudopotential approximation is somewhere between these two types of error: while an uncontrolled approximation in practice, it nevertheless targets a well-posed mathematical problem (the solution of a singular partial differential equation), and could in theory made systematically improvable.

The pseudopotential approximation has not been satisfactorily justified mathematically (although see [CM15] and [Dup17] for preliminary results). There are three distinct problems here: first, what is the error made in the frozen-core approximation, where core electrons are not modeled explicitly? Second, what is the error made by the pseudopotential approximation, which only tries to reproduce accurately the orbitals outside of a given cutoff radius? Third, how do the properties of the pseudopotential impact the efficiency of the discretization method? With Eric Cancès and Gaspard Kemlin, we hope to start addressing at least the first and third question.

VI.1.3. High-temperature regime. The behavior of electrons in the extreme conditions found in the center of planets or inertial confinement fusion has attracted much interest over the last years (see [PJ15] for an introduction). KSDFT is believed to degenerate to Thomas-Fermi theory in certain regimes of temperature and pressure, but the exact regime is still unclear, and would benefit from a rigorous mathematical analysis. Numerically, these systems are very challenging for KSDFT, requiring a large number of partially occupied orbitals as well as an adequate representation of high energies. There is a large scope for improved numerical methods. I plan to work on this, in collaboration with Marc Torrent (CEA).

VI.1.4. Computing with subspaces. Fundamentally, the unknowns in the zero-temperature Kohn-Sham equations are not the orbitals but rather the subspace they span. This is a theme that is gaining broader recognition in applied mathematics: the subspace is also the central object in dimensionality reduction, with applications to model reduction, computer vision and machine learning.

With Benjamin Stamm (Aachen), we plan to investigate systematically numerical computing with subspaces. In particular, notions of preconditioning, interpolation, and bifurcations, with clear geometrical interpretations in Euclidean space, become more subtle on the Grassmann manifold of subspaces. This has important consequences for numerical methods in electronic structure as well as other disciplines.

VI.2. Response properties

VI.2.1. Response properties of metals. The study presented in Sections III.3 and III.4 opens the way to a finer analysis of the properties of metals. First, due to the sharpness of the Fermi surface, their response to defects includes an oscillatory component, the Friedel oscillations [GV05]. This depends sensitively on the shape of the Fermi surface,

which itself depends on the composition of the material in a non-trivial way. A precise analysis of this dependence would greatly improve the mathematical understanding of the response properties of metals.

As we have shown in Section III.4, the response of metals to a uniform electric field is ballistic. To observe a finite conductivity, one needs to model collisions, either with impurities, lattice motion or other electrons. Mathematically, this would take the form of a Lindblad-type master equation, where collisions appear as dissipative terms, and whose semiclassical limit would be a Boltzmann-type equation. The derivation of these from first principles is a mathematical challenge (see [Spo06] for a related problem). Numerically, this has only started being possible in recent years, and has spurred a large activity in computational physics [Giu17].

VI.2.2. Time-dependent response. The time-dependent response of molecules or materials can be described in the time-dependent density functional theory (TDDFT) framework. In particular, linear response allows the derivation of spectra that can directly be compared to experiment. The efficient computation of these spectra is plagued by a fundamental difficulty: the finite systems used for numerical simulation necessarily have a discrete set of eigenvalues, whose approximation of the continuous spectrum is not at all clear. A number of numerical tricks are used to make the computations feasible: a small broadening parameter (corresponding to an absorption) is added, or complex absorbing potentials act to limit the propagation of waves to the inside of the computational domain [MPNE04; ALT17]. These methods often involve manual tuning, and a fully general solution procedure is desirable.

First, with Mi-Song Dupuy and Sören Behr (TU Münich), we are investigating the numerical analysis of the computation of response functions. In which regime, in which sense, and with which speed does the response function of the finite system converge to that of the full one? We are in particular aiming to prove pointwise convergence of the spectrum in the regime where $\frac{1}{L} \ll \eta \ll 1$, where L is the characteristic length of the computational domain, and η the broadening parameter.

A longer-term research project is to devise a fully automatic procedure to compute response functions, without manual selection of an appropriate dissipation mechanism and parameters. A source of inspiration could be the similar problems encountered in wave scattering in electromagnetism [BBDFT18].

VI.3. DFTK, the density functional toolkit

At the more practical level, the programs solving the Kohn-Sham equations are usually large codebases, developed over a large time period and optimized for efficiency. This makes it hard to develop new features, and leads to a disconnect between the numerical practices of different communities: mathematical physicists implement toy programs for 1D simplified models, numerical analysts test their method on artificial examples, and high-performance computing experts re-develop the part of the program they want to work on outside of the existing code bases. We are actively investigating the use of modern software development to solve this problem, utilizing in particular the newly-developed Julia language [BEKS17] to solve the "two-language problem" (where one prototypes in a language and implements in another). This will enable us to investigate the use of

methodology developed in part for machine learning applications (automatic differentiation, automatic parallelization, mixed-precision...) in the context of scientific computing. This is done in collaboration with Michael Herbst, who started as a postdoc in CERMICS/Inria in January 2019.

At the time of writing, the program, available freely at https://github.com/JuliaMolSim/DFTK.jl/, is able to perform ground-state computations of realistic materials (LDA/GGA DFT, with Goedecker pseudopotentials) with performance comparable to widely used packages, while remaining relatively simple (about $3,000$ lines of code). It is actively used for research in numerical methods in the group at CERMICS/Inria, and is intended to evolve into a code able to compute solutions of the Kohn-Sham equations with guaranteed error bounds.

List of papers

This is a list of all my papers, including those published during my PhD thesis, in reverse chronologic order.

[1] E. Cancès, C. Fermanian Kammerer, A. Levitt, S. Siraj-Dine. Coherent electronic transport in periodic crystals. *Submitted*, 2020.

[2] A. Levitt. Screening in the finite-temperature reduced Hartree-Fock model. Accepted in *Archive for Rational Mechanics and Applications*, 2018.

[3] E. Cancès, V. Ehrlacher, D. Gontier, A. Levitt, and D. Lombardi. Numerical quadrature in the Brillouin zone for periodic Schrödinger operators. Accepted in *Numerische Mathematik*, 2019.

[4] D. Gontier, A. Levitt, and S. Siraj-Dine. Numerical construction of Wannier functions through homotopy. *Journal of Mathematical Physics*, 60(3):031901, 2019.

[5] A. Damle, A. Levitt, and L. Lin. Variational formulation for Wannier functions with entangled band structure. *Multiscale Modeling & Simulation*, 17(1):167–191, 2019.

[6] H.D. Cornean, D. Gontier, A. Levitt, and D. Monaco. Localised Wannier functions in metallic systems. In *Annales Henri Poincaré*, pages 1–25. Springer, 2017.

[7] E. Cancès, A. Levitt, G. Panati, and G. Stoltz. Robust determination of maximally localized Wannier functions. *Physical Review B*, 95(7):075114, 2017.

[8] X. Antoine, A. Levitt, and Q. Tang. Efficient spectral computation of the stationary states of rotating Bose–Einstein condensates by preconditioned nonlinear conjugate gradient methods. *Journal of Computational Physics*, 343:92–109, 2017.

[9] A. Levitt and C. Ortner. Convergence and cycling in walker-type saddle search algorithms. *SIAM Journal on Numerical Analysis*, 55(5):2204–2227, 2017.

[10] F. Aviat, A. Levitt, B. Stamm, Y. Maday, P. Ren, J.W. Ponder, L. Lagardere, and J-P. Piquemal. Truncated conjugate gradient: an optimal strategy for the analytical evaluation of the many-body polarization energy and forces in molecular simulations. *Journal of chemical theory and computation*, 13(1):180–190, 2016.

[11] X. Gonze, F. Jollet, et al. Recent developments in the ABINIT software package. *Computer Physics Communications*, 205:106–131, 2016.

[12] A. Levitt and M. Torrent. Parallel eigensolvers in plane-wave density functional theory. *Computer Physics Communications*, 187:98–105, 2015.

[13] A. Levitt. Solutions of the multiconfiguration Dirac–Fock equations. *Reviews in Mathematical Physics*, 26(07), 2014.

[14] A. Levitt. Best constants in Lieb-Thirring inequalities: a numerical investigation. *Journal of Spectral Theory*, 4(1):153–175, 2014.

[15] A. Levitt. Convergence of gradient-based algorithms for the Hartree-Fock equations." *ESAIM: Mathematical Modelling and Numerical Analysis*, 46(6):1321–1336, 2012.

[16] G. James, A. Levitt, C. Ferreira. Continuation of discrete breathers from infinity in a nonlinear model for DNA breathing. *Applicable Analysis*, 89(9):1447-1465, 2010.

Bibliography

[AC09] A. Anantharaman and E. Cancès. "Existence of minimizers for Kohn–Sham models in quantum chemistry". *Annales de l'Institut Henri Poincare (C) Non Linear Analysis* 26.6 (2009), pp. 2425–2455.

[Adl62] S. L. Adler. "Quantum theory of the dielectric constant in real solids". *Physical Review* 126.2 (1962), p. 413.

[AE06] J.-P. Aubin and I. Ekeland. *Applied nonlinear analysis*. Courier Corporation, 2006.

[AG98] F. Aryasetiawan and O. Gunnarsson. "The GW method". *Reports on Progress in Physics* 61.3 (1998), p. 237.

[ALT17] X. Antoine, E. Lorin, and Q. Tang. "A friendly review of absorbing boundary conditions and perfectly matched layers for classical and relativistic quantum waves equations". *Molecular Physics* 115.15-16 (2017), pp. 1861–1879.

[AMS09] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[And65] D. G. Anderson. "Iterative procedures for nonlinear integral equations". *Journal of the ACM (JACM)* 12.4 (1965), pp. 547–560.

[Bae06] M. Baer. *Beyond Born-Oppenheimer: electronic nonadiabatic coupling terms and conical intersections*. John Wiley & Sons, 2006.

[BBDFT18] A.-S. Bonnet-Ben Dhia, S. Fliss, and A. Tonnoir. "The halfspace matching method: A new method to solve scattering problems in infinite media". *Journal of Computational and Applied Mathematics* 338 (2018), pp. 44–68.

[BBIMJNP+09] B. R. Brooks, C. L. Brooks III, A. D. Mackerell Jr, L. Nilsson, R. J. Petrella, et al. "CHARMM: the biomolecular simulation program". *Journal of computational chemistry* 30.10 (2009), pp. 1545–1614.

[BBL81] R. Benguria, H. Brézis, and E. H. Lieb. "The Thomas-Fermi-von Weizsäcker theory of atoms and molecules". *Communications in Mathematical Physics* 79.2 (1981), pp. 167–180.

[BC12] W. Bao and Y. Cai. "Mathematical theory and numerical methods for Bose-Einstein condensation". *Kinetic & Related Models* 6 (2012).

[BDDRV00] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst. *Templates for the solution of algebraic eigenvalue problems: a practical guide*. SIAM, 2000.

[BDGDCG01] S. Baroni, S. De Gironcoli, A. Dal Corso, and P. Giannozzi. "Phonons and related crystal properties from density-functional perturbation theory". *Reviews of Modern Physics* 73.2 (2001), p. 515.

[Bec93] A. D. Becke. "Density-functional thermochemistry. III. The role of exact exchange". *The Journal of chemical physics* 98.7 (1993), pp. 5648–5652.

[BEKS17] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. "Julia: A fresh approach to numerical computing". *SIAM review* 59.1 (2017), pp. 65–98.

[BEP97] K. Burke, M. Ernzerhof, and J. P. Perdew. “The adiabatic connection method: a non-empirical hybrid”. *Chemical Physics Letters* 265.1-2 (1997), pp. 115–120.

[Ber84] M. V. Berry. “Quantal phase factors accompanying adiabatic changes”. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences* 392.1802 (1984), pp. 45–57.

[BESB94] J. Bellissard, A. van Elst, and H. Schulz-Baldes. “The noncommutative geometry of the quantum Hall effect”. *Journal of Mathematical Physics* 35.10 (1994), pp. 5373–5451.

[BGKS05] J.-M. Bouclet, F. Germinet, A. Klein, and J. H. Schenker. “Linear response theory for magnetic Schrödinger operators in disordered media”. *Journal of Functional Analysis* 226.2 (2005), pp. 301–372.

[BJA94] P. E. Blöchl, O. Jepsen, and O. K. Andersen. “Improved tetrahedron method for Brillouin-zone integrations”. *Physical Review B* 49.23 (1994), p. 16223.

[Blö94] P. E. Blöchl. “Projector augmented-wave method”. *Physical review B* 50.24 (1994), p. 17953.

[BPCMM07] C. Brouder, G. Panati, M. Calandra, C. Mourougane, and N. Marzari. “Exponential localization of Wannier functions in insulators”. *Physical review letters* 98.4 (2007), p. 046402.

[BW13] K. Burke and L. O. Wagner. “DFT in a nutshell”. *International Journal of Quantum Chemistry* 113.2 (2013), pp. 96–101.

[BWG05] K. Burke, J. Werschnik, and E. Gross. “Time-dependent density functional theory: Past, present, and future”. *The Journal of chemical physics* 123.6 (2005), p. 062206.

[Can01] E. Cancès. “Self-consistent field algorithms for Kohn–Sham models with fractional occupation numbers”. *The Journal of Chemical Physics* 114.24 (2001), pp. 10616–10622.

[CCIDGL+05] D. A. Case, T. E. Cheatham III, T. Darden, H. Gohlke, R. Luo, et al. “The Amber biomolecular simulation programs”. *Journal of computational chemistry* 26.16 (2005), pp. 1668–1688.

[CCM12] E. Cancès, R. Chakir, and Y. Maday. “Numerical analysis of the planewave discretization of some orbital-free and Kohn-Sham models”. *ESAIM: Mathematical Modelling and Numerical Analysis* 46.2 (2012), pp. 341–388.

[CDL08] E. Cancès, A. Deleurence, and M. Lewin. “A new approach to the modeling of local defects in crystals: The reduced Hartree-Fock case”. *Communications in Mathematical Physics* 281.1 (2008), pp. 129–177.

[CDMSV14] E. Cancès, G. Dusson, Y. Maday, B. Stamm, and M. Vohralík. “A perturbation-method-based a posteriori estimator for the planewave discretization of nonlinear Schrödinger equations”. *Comptes Rendus Mathématique* 352.11 (2014), pp. 941–946.

[CJL06] E. Cancès, B. Jourdain, and T. Lelievre. "Quantum Monte Carlo simulations of fermions: a mathematical analysis of the fixed-node approximation". *Mathematical Models and Methods in Applied Sciences* 16.09 (2006), pp. 1403–1440.

[CKST03] E. Cances, K. N. Kudin, G. E. Scuseria, and G. Turinici. "Quadratically convergent algorithm for fractional occupation numbers in density functional theory". *The Journal of chemical physics* 118.12 (2003), pp. 5364–5368.

[CL00] P. M. Chaikin and T. C. Lubensky. *Principles of condensed matter physics*. Vol. 1. Cambridge university press Cambridge, 2000.

[CL10] É. Cancès and M. Lewin. "The dielectric permittivity of crystals in the reduced Hartree–Fock approximation". *Archive for Rational Mechanics and Analysis* 197.1 (2010), pp. 139–177.

[CLB00a] E. Cancès and C. Le Bris. "Can we outperform the DIIS approach for electronic structure calculations?" *International Journal of Quantum Chemistry* 79.2 (2000), pp. 82–90.

[CLB00b] E. Cancès and C. Le Bris. "On the convergence of SCF algorithms for the Hartree-Fock equations". *ESAIM: Mathematical Modelling and Numerical Analysis* 34.4 (2000), pp. 749–774.

[CLB13] E. Cances and C. Le Bris. "Mathematical modeling of point defects in materials science". *Mathematical Models and Methods in Applied Sciences* 23.10 (2013), pp. 1795–1859.

[CLBL01] I. Catto, C. Le Bris, and P.-L. Lions. "On the thermodynamic limit for Hartree–Fock type models". *Annales de l'Institut Henri Poincaré (C) Non Linear Analysis* 18.6 (2001), pp. 687–760.

[CM15] E. Cancès and N. Mourad. "Existence of a type of optimal norm-conserving pseudopotentials for Kohn–Sham models". *Communications in Mathematical Sciences* 14.5 (2015), pp. 1315–1352.

[DG12] R. M. Dreizler and E. K. Gross. *Density functional theory: an approach to the quantum many-body problem*. Springer Science & Business Media, 2012.

[Dir29] P. A. M. Dirac. "Quantum mechanics of many-electron systems". *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 123.792 (1929), pp. 714–733.

[DLY15] A. Damle, L. Lin, and L. Ying. "Compressed representation of Kohn–Sham orbitals via selected columns of the density matrix". *Journal of chemical theory and computation* 11.4 (2015), pp. 1463–1469.

[DM17] G. Dusson and Y. Maday. "A posteriori analysis of a nonlinear Gross–Pitaevskii-type eigenvalue problem". *IMA Journal of Numerical Analysis* 37.1 (2017), pp. 94–137.

[Dol+00] M. Dolg et al. "Effective core potentials". *Modern methods and algorithms of quantum chemistry* 1 (2000), pp. 479–508.

[DSYG18] J. A. Duersch, M. Shao, C. Yang, and M. Gu. "A robust and efficient implementation of LOBPCG". *SIAM Journal on Scientific Computing* 40.5 (2018), pp. C655–C676.

[Dup17] M.-S. Dupuy. “Projector augmented-wave method: an analysis in a one-dimensional setting”. *arXiv preprint arXiv:1712.04685* (2017).

[EAS98] A. Edelman, T. A. Arias, and S. T. Smith. “The geometry of algorithms with orthogonality constraints”. *SIAM journal on Matrix Analysis and Applications* 20.2 (1998), pp. 303–353.

[EC59] H Ehrenreich and M. H. Cohen. “Self-consistent field approach to the many-electron problem”. *Physical Review* 115.4 (1959), p. 786.

[ERVE02] W. E, W. Ren, and E. Vanden-Eijnden. “String method for the study of rare events”. *Physical Review B* 66.5 (2002), p. 052301.

[FB60] J. Foster and S. Boys. “Canonical configurational interaction procedure”. *Reviews of Modern Physics* 32.2 (1960), p. 300.

[FBN09] C. Freysoldt, S. Boeck, and J. Neugebauer. “Direct minimization technique for metals in density functional theory”. *Physical Review B* 79.24 (2009), p. 241103.

[Fef85] C. Fefferman. “The thermodynamic limit for a crystal”. *Communications in mathematical physics* 98.3 (1985), pp. 289–311.

[FMP16] D. Fiorenza, D. Monaco, and G. Panati. “Construction of real-valued localized composite Wannier functions for insulators”. *Annales Henri Poincaré* 17.1 (2016), pp. 63–97.

[FW12] C. Fefferman and M. Weinstein. “Honeycomb lattice potentials and Dirac points”. *Journal of the American Mathematical Society* 25.4 (2012), pp. 1169–1220.

[Geo04] A. Georges. “Strongly Correlated Electron Materials: Dynamical Mean-Field Theory and Electronic Structure”. *AIP Conference Proceedings*. Vol. 715. 1. AIP. 2004, pp. 3–74.

[Gia10] P. Giannozzi. *Notes on pseudopotential generation*. https://www.quantum-espresso.org/Doc/pseudo-gen.pdf. 2010.

[Giu17] F. Giustino. “Electron-phonon interactions from first principles”. *Reviews of Modern Physics* 89.1 (2017), p. 015003.

[GL16] D. Gontier and S. Lahbabi. “Convergence rates of supercell calculations in the reduced Hartree- Fock model”. *ESAIM: Mathematical Modelling and Numerical Analysis* 50.5 (2016), pp. 1403–1424.

[GOP16] N Gould, C. Ortner, and D Packwood. “A dimer-type saddle search algorithm with preconditioning and linesearch”. *Mathematics of Computation* 85.302 (2016), pp. 2939–2966.

[Gri+89] A. Griewank et al. “On automatic differentiation”. *Mathematical Programming: recent developments and applications* 6.6 (1989), pp. 83–107.

[Gri11] S. Grimme. “Density functional theory with London dispersion corrections”. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1.2 (2011), pp. 211–228.

[GTH96] S Goedecker, M Teter, and J. Hutter. “Separable dual-space Gaussian pseudopotentials”. *Physical Review B* 54.3 (1996), p. 1703.

[GV05] G. Giuliani and G. Vignale. *Quantum theory of the electron liquid*. Cambridge university press, 2005.

[Hal88] F. D. M. Haldane. “Model for a quantum Hall effect without Landau levels: Condensed-matter realization of the" parity anomaly"”. *Physical review letters* 61.18 (1988), p. 2015.

[Ham13] D. Hamann. “Optimized norm-conserving Vanderbilt pseudopotentials”. *Physical Review B* 88.8 (2013), p. 085117.

[HGH98] C Hartwigsen, S. Gœdecker, and J. Hutter. “Relativistic separable dual-space Gaussian pseudopotentials from H to Rn”. *Physical Review B* 58.7 (1998), p. 3641.

[HJ00] G. A. Hagedorn and A. Joye. “Exponentially accurate semiclassical dynamics: propagation, localization, Ehrenfest times, scattering, and more general states”. *Annales Henri Poincaré* 1.5 (2000), pp. 837–883.

[HJO14] T. Helgaker, P. Jorgensen, and J. Olsen. *Molecular electronic-structure theory*. John Wiley & Sons, 2014.

[HK64] P. Hohenberg and W. Kohn. “Inhomogeneous electron gas”. *Physical review* 136.3B (1964), B864.

[HL06] U. Hetmaniuk and R. Lehoucq. “Basis selection in LOBPCG”. *Journal of Computational Physics* 218.1 (2006), pp. 324–332.

[HLR94] B. L. Hammond, W. A. Lester, and P. J. Reynolds. *Monte Carlo methods in ab initio quantum chemistry*. Vol. 1. World Scientific, 1994.

[HS12] P. D. Hislop and I. M. Sigal. *Introduction to spectral theory: With applications to Schrödinger operators*. Vol. 113. Springer Science & Business Media, 2012.

[Hum79] C. Humphreys. “STEM imaging of crystals and defects”. *Introduction to analytical electron microscopy*. Springer, 1979, pp. 305–332.

[JDRI13] L. Joubert-Doriol, I. G. Ryabinkin, and A. F. Izmaylov. “Geometric phase effects in low-energy dynamics near conical intersections: A study of the multidimensional linear vibronic coupling model”. *The Journal of chemical physics* 139.23 (2013), p. 234103.

[JMJ98] H. Jónsson, G. Mills, and K. W. Jacobsen. “Nudged elastic band method for finding minimum energy paths of transitions” (1998).

[JOHCR+13] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, et al. “Commentary: The Materials Project: A materials genome approach to accelerating materials innovation”. *Apl Materials* 1.1 (2013), p. 011002.

[Kat51] T. Kato. “Fundamental properties of Hamiltonian operators of Schrödinger type”. *Transactions of the American Mathematical Society* 70.2 (1951), pp. 195–211.

[KB82] L. Kleinman and D. Bylander. “Efficacious form for model pseudopotentials”. *Physical Review Letters* 48.20 (1982), p. 1425.

[Ker81] G. Kerker. “Efficient iteration scheme for self-consistent pseudopotential calculations”. *Physical Review B* 23.6 (1981), p. 3082.

[KM05] C. L. Kane and E. J. Mele. “$\mathbb{Z}_2$ topological order and the quantum spin Hall effect”. *Physical review letters* 95.14 (2005), p. 146802.

[Kny01] A. V. Knyazev. “Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method”. *SIAM journal on scientific computing* 23.2 (2001), pp. 517–541.

[Kny98] A. V. Knyazev. “Preconditioned eigensolvers—an oxymoron”. *Electron. Trans. Numer. Anal* 7 (1998), pp. 104–123.

[KS65] W. Kohn and L. J. Sham. “Self-consistent equations including exchange and correlation effects”. *Physical review* 140.4A (1965), A1133.

[Kub57] R. Kubo. “Statistical-mechanical theory of irreversible processes. I. General theory and simple applications to magnetic and conduction problems”. *Journal of the Physical Society of Japan* 12.6 (1957), pp. 570–586.

[LB93] C. Le Bris. “Quelques problèmes mathématiques en chimie quantique moléculaire”. PhD thesis. Ecole polytechnique, 1993.

[LBBBB+16] K. Lejaeghere, G. Bihlmayer, T. Björkman, P. Blaha, S. Blügel, et al. “Reproducibility in density functional theory calculations of solids”. *Science* 351.6280 (2016), aad3000.

[LBL05] C. Le Bris and P.-L. Lions. “From atoms to crystals: a mathematical journey”. *Bulletin of the American Mathematical Society* 42.3 (2005), pp. 291–363.

[Lev79] M. Levy. “Universal variational functionals of electron densities, first-order density matrices, and natural spin-orbitals and solution of the v-representability problem”. *Proceedings of the National Academy of Sciences* 76.12 (1979), pp. 6062–6065.

[Lie83] E. H. Lieb. “Density functionals for coulomb systems”. *International Journal of Quantum Chemistry* 24.3 (1983), pp. 243–277.

[Lin54] J. Lindhard. “On the properties of a gas of charged particles”. *Dan. Vid. Selsk Mat.-Fys. Medd.* 28 (1954), p. 8.

[Lio87] P.-L. Lions. “Solutions of Hartree-Fock equations for Coulomb systems”. *Communications in Mathematical Physics* 109.1 (1987), pp. 33–97.

[LLS19] M. Lewin, E. H. Lieb, and R. Seiringer. “The Local Density Approximation in Density Functional Theory”. *arXiv preprint arXiv:1903.04046* (2019).

[LS77a] E. H. Lieb and B. Simon. “The hartree-fock theory for coulomb systems”. *Communications in Mathematical Physics* 53.3 (1977), pp. 185–194.

[LS77b] E. H. Lieb and B. Simon. “The Thomas-Fermi theory of atoms, molecules and solids”. *Advances in mathematics* 23.1 (1977), pp. 22–116.

[LY13] L. Lin and C. Yang. “Elliptic preconditioner for accelerating the self-consistent field iteration in Kohn–Sham density functional theory”. *SIAM Journal on Scientific Computing* 35.5 (2013), S277–S298.

[MCCL15] J. I. Mustafa, S. Coh, M. L. Cohen, and S. G. Louie. “Automated construction of maximally localized Wannier functions: Optimized projection functions method”. *Physical Review B* 92.16 (2015), p. 165134.

[MMYSV12] N. Marzari, A. A. Mostofi, J. R. Yates, I. Souza, and D. Vanderbilt. “Maximally localized Wannier functions: Theory and applications”. *Reviews of Modern Physics* 84.4 (2012), p. 1419.

[MP89] M. Methfessel and A. Paxton. “High-precision sampling for Brillouin-zone integration in metals”. *Physical Review B* 40.6 (1989), p. 3616.

[MPNE04] J. Muga, J. Palao, B Navarro, and I. Egusquiza. "Complex absorbing potentials". *Physics Reports* 395.6 (2004), pp. 357–426.

[MS14] J. Málek and Z. Strakos. *Preconditioning and the conjugate gradient method in the context of solving PDEs*. Vol. 1. SIAM, 2014.

[MV97] N. Marzari and D. Vanderbilt. "Maximally localized generalized Wannier functions for composite energy bands". *Physical review B* 56.20 (1997), p. 12847.

[Neu09] J. Neuberger. *Sobolev gradients and differential equations*. Springer Science & Business Media, 2009.

[Nie93] F. Nier. "A variational formulation of Schrödinger-Poisson systems in dimension $d \leq 3$". *Communications in partial differential equations* 18.7-8 (1993), pp. 1125–1147.

[NRS18] P. Norman, K. Ruud, and T. Saue. *Principles and practices of molecular properties: Theory, modeling, and simulations*. John Wiley & Sons, 2018.

[NW06] J. Nocedal and S. Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[NW29] J. von Neumann and E. Wigner. "On the behaviour of eigenvalues in adiabatic processes". *Phys. Z.* 30 (1929).

[Pan07] G. Panati. "Triviality of Bloch and Bloch–Dirac bundles". *Annales Henri Poincaré* 8.5 (2007), pp. 995–1011.

[PBE96] J. P. Perdew, K. Burke, and M. Ernzerhof. "Generalized gradient approximation made simple". *Physical review letters* 77.18 (1996), p. 3865.

[PCSMK16] G. Pizzi, A. Cepellotti, R. Sabatini, N. Marzari, and B. Kozinsky. "AiiDA: automated interactive infrastructure and database for computational science". *Computational Materials Science* 111 (2016), pp. 218–230.

[PJ15] A. Pribram-Jones. "Thermal density functional theory, ensemble density functional theory, and potential functional theory for warm dense matter". PhD thesis. UC Irvine, 2015.

[PM01] C. J. Pickard and F. Mauri. "All-electron magnetic response with pseudopotentials: NMR chemical shifts". *Physical Review B* 63.24 (2001), p. 245101.

[PP13] G. Panati and A. Pisante. "Bloch bundles, Marzari-Vanderbilt functional and maximally localized Wannier functions". *Communications in Mathematical Physics* 322.3 (2013), pp. 835–875.

[PP99] C. Pickard and M. Payne. "Extrapolative approaches to Brillouin-zone integration". *Physical Review B* 59.7 (1999), p. 4685.

[PS08] C. J. Pethick and H. Smith. *Bose–Einstein condensation in dilute gases*. Cambridge university press, 2008.

[PST03] G. Panati, H. Spohn, and S. Teufel. "Effective dynamics for Bloch electrons: Peierls substitution and beyond". *Communications in Mathematical Physics* 242.3 (2003), pp. 547–578.

[PTAAJ92] M. C. Payne, M. P. Teter, D. C. Allan, T. Arias, and a. J. Joannopoulos. “Iterative minimization techniques for ab initio total-energy calculations: molecular dynamics and conjugate gradients”. *Reviews of Modern Physics* 64.4 (1992), p. 1045.

[Pul82] P. Pulay. “Improved SCF convergence acceleration”. *Journal of Computational Chemistry* 3.4 (1982), pp. 556–560.

[PW92] J. P. Perdew and Y. Wang. “Accurate and simple analytic representation of the electron-gas correlation energy”. *Physical Review B* 45.23 (1992), p. 13244.

[PWRPC+10] J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, et al. “Current status of the AMOEBA polarizable force field”. *The Journal of Physical Chemistry B* 114.8 (2010), pp. 2549–2564.

[Pyy12] P. Pyykkö. “Relativistic effects in chemistry: more common than you thought”. *Annual Review of Physical Chemistry* 63 (2012), pp. 45–64.

[RRKJ90] A. M. Rappe, K. M. Rabe, E. Kaxiras, and J. Joannopoulos. “Optimized pseudopotentials”. *Physical Review B* 41.2 (1990), p. 1227.

[RS11] T. Rohwedder and R. Schneider. “An analysis for the DIIS acceleration method used in quantum chemistry calculations”. *Journal of Mathematical Chemistry* 49.9 (2011), p. 1889.

[RS72] M. Reed and B. Simon. *Methods of modern mathematical physics. I, II, III, IV*. Academic Press Inc., New York, 1972.

[Rut70] H. Rutishauser. “Simultaneous iteration method for symmetric matrices”. *Numerische Mathematik* 16.3 (1970), pp. 205–223.

[RV07] R. Resta and D. Vanderbilt. “Theory of polarization: a modern approach”. *Physics of Ferroelectrics*. Springer, 2007, pp. 31–68.

[RZK66] L. M. Roth, H. Zeiger, and T. Kaplan. “Generalization of the Ruderman-Kittel-Kasuya-Yosida interaction for nonspherical Fermi surfaces”. *Physical Review* 149.2 (1966), p. 519.

[Saa11] Y. Saad. *Numerical methods for large eigenvalue problems: revised edition*. Vol. 66. Siam, 2011.

[Shi96] E. L. Shirley. “Optimal basis sets for detailed Brillouin-zone integrations”. *Physical Review B* 54.23 (1996), p. 16464.

[Sim00] B. Simon. “Schrödinger operators in the twentieth century”. *Journal of Mathematical Physics* 41.6 (2000), pp. 3523–3555.

[Sim83] B. Simon. “Holonomy, the quantum adiabatic theorem, and Berry’s phase”. *Physical Review Letters* 51.24 (1983), p. 2167.

[SMV01] I. Souza, N. Marzari, and D. Vanderbilt. “Maximally localized Wannier functions for entangled energy bands”. *Physical Review B* 65.3 (2001), p. 035109.

[SPISCIB15] A. C. Simmonett, F. C. Pickard IV, Y. Shao, T. E. Cheatham III, and B. R. Brooks. “Efficient treatment of induced dipoles”. *The Journal of chemical physics* 143.7 (2015), p. 074115.

[Spo06] H. Spohn. “The phonon Boltzmann equation, properties and link to weakly anharmonic lattice dynamics”. *Journal of statistical physics* 124.2-4 (2006), pp. 1041–1104.

[SSB18] J. C. Smith, F. Sagredo, and K. Burke. "Warming up density functional theory". *Frontiers of Quantum Chemistry*. Springer, 2018, pp. 249–271.

[TBI97] L. N. Trefethen and D. Bau III. *Numerical linear algebra*. Vol. 50. SIAM, 1997.

[Teu03] S. Teufel. *Adiabatic perturbation theory in quantum dynamics*. Springer, 2003.

[Tho81] B. T. Thole. "Molecular polarizabilities calculated with a modified dipole interaction". *Chemical Physics* 59.3 (1981), pp. 341–350.

[TKNN82] D. J. Thouless, M. Kohmoto, M. P. Nightingale, and M. den Nijs. "Quantized Hall conductance in a two-dimensional periodic potential". *Physical review letters* 49.6 (1982), p. 405.

[TM91] N. Troullier and J. L. Martins. "Efficient pseudopotentials for plane-wave calculations". *Physical review B* 43.3 (1991), p. 1993.

[TW14] L. N. Trefethen and J. Weideman. "The exponentially convergent trapezoidal rule". *SIAM Review* 56.3 (2014), pp. 385–458.

[VPMYM+19] V. Vitale, G. Pizzi, A. Marrazzo, J. Yates, N. Marzari, et al. "Automated high-throughput wannierisation". *arXiv preprint arXiv:1909.00433* (2019).

[Wan37] G. H. Wannier. "The structure of electronic excitation levels in insulating crystals". *Physical Review* 52.3 (1937), p. 191.

[Wis63] N. Wiser. "Dielectric constant with local field effects included". *Physical Review* 129.1 (1963), p. 62.

[WN11] H. F. Walker and P. Ni. "Anderson acceleration for fixed-point iterations". *SIAM Journal on Numerical Analysis* 49.4 (2011), pp. 1715–1735.

[WS05] W. Wang and R. D. Skeel. "Fast evaluation of polarizable forces". *The Journal of chemical physics* 123.16 (2005), p. 164107.

[YWVS07] J. R. Yates, X. Wang, D. Vanderbilt, and I. Souza. "Spectral and Fermi surface properties from Wannier interpolation". *Physical Review B* 75.19 (2007), p. 195121.

[ZSTC06] Y. Zhou, Y. Saad, M. L. Tiago, and J. R. Chelikowsky. "Self-consistent-field calculations using Chebyshev-filtered subspace iteration". *Journal of Computational Physics* 219.1 (2006), pp. 172–184.