

Université Marie et Louis Pasteur

Habilitation à diriger des recherches

présentée par

Geneviève Dusson

Mathematical and numerical analysis in quantum chemistry and materials science

---

Soutenue publiquement le 16 décembre 2025 devant le jury composé de :

Mme	Julie Delon	Examinatrice
M.	Alexei Lozinski	Examinateur
M.	Jianfeng Lu	Rapporteur
M.	Yvon Maday	Examinateur
M.	Mihai-Cosmin Marinica	Rapporteur
M.	Anthony Nouy	Rapporteur
M.	Julien Toulouse	Examinateur

Remerciements

Je tiens tout d'abord à remercier chaleureusement les rapporteurs, Jianfeng Lu, Cosmin Marinica et Anthony Nouy, pour le temps qu'ils ont consacré à mon habilitation, ainsi que tous les membres du jury d'avoir accepté de participer à cette soutenance : Julie Delon, Alexei Lozinski, Yvon Maday, et Julien Toulouse.

Merci à toutes les personnes qui m'ont soutenues dans l'aventure de l'HDR, en particulier Virginie Ehrlacher et Antoine Levitt, pour les nombreux conseils sur le manuscrit, et Yvon Maday, pour m'avoir régulièrement relancée pour écrire et déposer le manuscrit.

Au cours de ces années à Jussieu, Warwick et Besançon, j'ai eu l'opportunité de rencontrer et d'être accompagnée par des personnes aux qualités humaines et scientifiques remarquables, qui ont largement contribué à mon parcours. Il serait trop long de citer ici l'ensemble des personnes concernées, mais je souhaite néanmoins remercier plus particulièrement certaines d'entre elles.

Tout d'abord, merci infiniment à Yvon Maday, Eric Cancès, Benjamin Stamm, et Martin Vohralík pour m'avoir formée pendant la thèse, et pour leur soutien constant depuis.

Merci également à Christoph Ortner pour son accueil formidable en postdoc ainsi que toutes les personnes rencontrées à Warwick autour des potentiels interatomiques.

Un grand merci à tous les bisontin·e·s, notamment les membres du LmB pour leur accueil et l'ambiance chaleureuse du laboratoire, qui en fait un lieu de travail très agréable. Je remercie sincèrement les différents services techniques du laboratoire pour leur soutien sans faille, sans qui nous ne pourrions travailler.

Je souhaite ensuite exprimer ma gratitude aux membres de l'équipe Matherials pour l'accueil précieux qu'ils m'ont réservé au Cermics, et en particulier, parmi ceux que je n'ai pas encore remerciés, Tony Lelièvre et Gabriel Stoltz.

J'ai également une pensée pour toutes ces semaines passées à Roscoff, où les discussions scientifiques intenses et l'air de la mer forment une symbiose toujours très productive et stimulante. Merci à toutes celles et ceux avec qui j'ai pu avoir des discussions scientifiques là-bas.

Merci enfin à toutes mes collaboratrices et collaborateurs, ainsi qu'aux étudiantes et étudiants, qui me font confiance. Leur dynamisme, leurs questionnements et leur implication dans nos projets communs constituent un moteur essentiel de ma réflexion scientifique.

Ce manuscrit est dédié à ma famille pour son soutien indéfectible.

Contents

Chapter 1

Introduction

My research so far focuses on the mathematical and numerical analysis of models and numerical methods designed for the simulation of molecules and materials systems. In this chapter, I first introduce the main models that are analyzed in the subsequent chapters. I then present scientific questions motivating this research. Finally, I provide a brief description of my contributions to this field together with a list of publications.

1.1 Models in molecular simulation

Numerical simulations performed for computing physical properties of molecular and materials systems represent a large part of calculations run on supercomputers (around 35% for the Swiss supercomputers in 2021 [1]). Depending on the size of the considered system, the available computational resources, and the properties that are computed, various methods can be employed, each providing a different level of accuracy.

Electronic structure (or *ab initio*) methods rely on the explicit modeling of the electrons in the system. These methods include e.g. wave-function methods and density functional theory methods. With the current computational resources available they can typically simulate moderate-size molecular systems with only up to a few hundreds atoms. Mathematically, the electronic structure methods rely on solving partial differential equations, which explains why they are computationally demanding. As a main advantage, they are transferable: from one physical system to the next, only the parameters of the partial differential equations change. These parameters, moreover, only consist of masses and charges of the atoms in the systems, as well as a few fundamental constants of physics.

For larger systems, classical methods such as interatomic potentials and coarse-grained models can be used. In this case, the modeling of the electrons is implicit, and the physical properties, starting from the potential energy of the system are written as explicit functions of the nuclei positions. These functions are in general fitted to match either experimental data or *ab initio* data. As an advantage these methods can simulate systems with at least millions to billions of atoms, but in general lack transferability: they are often inaccurate outside the chemical environments they were fitted or designed for.

The following Chapters 2 and 4 are concerned with electronic structure models, and Chapter 5 mostly deals with interatomic potentials. We therefore start by briefly describing the main features of electronic structure models in Section 1.1.1 before presenting interatomic potentials in Section 1.1.3.

1.1.1 Electronic structure calculations

In the context of electronic structure calculations, molecular systems are often described by classical nuclei characterized by their positions and velocities, and quantum electrons modeled by a wavefunction. Indeed, protons being heavier than electrons, electrons move faster than nuclei, and can be considered relaxed at each time step of the nuclei movement. This is a standard approximation in quantum chemistry called the Born–Oppenheimer approximation [30], which is considered throughout this manuscript. Thus we consider a molecular system with M nuclei and N_{el} electrons. The positions of the nuclei are denoted by $\mathbf{R} \in \mathbb{R}^{3M}$.

We are particularly interested in computing the ground state of the system, which is the state of lowest energy. This is a major problem in electronic structure calculations, as it gives access to the most stable state of the system, and is a required step to compute more involved properties, such as the characterization of excited states. The electronic wavefunction, for given nuclei positions $\mathbf{R}$, is denoted by $\Psi_{\mathbf{R}} : \mathbb{R}^{3N_{\mathrm{el}}} \to \mathbb{C}$ and is anti-symmetric with respect to permutations of the N_{el} input variables in $\mathbb{R}^3$. Since ground state wavefunctions can be chosen real-valued (if $\Psi_{\mathbf{R}}$ is a ground state, $\mathrm{Re}(\Psi_{\mathbf{R}})$ and $\mathrm{Im}(\Psi_{\mathbf{R}})$ are also ground states), we simplify the presentation accordingly and consider real-valued wavefunctions. The behavior of the electrons can then be modeled by the Schrödinger equation, a linear high-dimensional partial differential equation, which writes

$$H_{\mathbf{R}}\Psi_{\mathbf{R}} = E_{\mathbf{R}}\Psi_{\mathbf{R}}, \tag{1.1.1}$$

where $H_{\mathbf{R}}$ is the Hamiltonian of the problem, given in atomic units, and omitting the spin variables for simplicity, by

$$H_{\mathbf{R}} = -\frac{1}{2}\sum_{i=1}^{N_{\mathrm{el}}} \Delta_{\mathbf{r}_i} + \sum_{i=1}^{N_{\mathrm{el}}} V_{\mathbf{R}}^{\mathrm{ne}}(\mathbf{r}_i) + \sum_{1\le i<j\le N_{\mathrm{el}}} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|}, \quad V_{\mathbf{R}}^{\mathrm{ne}}(\mathbf{r}_i) = -\sum_{k=1}^{M} \frac{z_k}{|\mathbf{R}_k - \mathbf{r}_i|}.$$

It is composed of three parts respectively corresponding to (i) the kinetic part, (ii) the Coulomb interaction between the nuclei and the electrons, and (iii) the Coulomb interaction between the electrons. The parameters z_k are the charges of the nuclei, and $E_{\mathbf{R}}$ is the ground state energy of the system. The ground state energy also corresponds to the minimum of the following constrained minimization problem

$$\inf_{\substack{\Psi\in H^1_{\mathrm{as}}(\mathbb{R}^{3N_{\mathrm{el}}}) \\ \|\Psi\|_{L^2}=1}} \langle \Psi, H_{\mathbf{R}}\Psi\rangle, \tag{1.1.2}$$

where $H^1_{\mathrm{as}}(\mathbb{R}^{3N_{\mathrm{el}}})$ is the space of antisymmetric H^1 functions

$$H^1_{\mathrm{as}}(\mathbb{R}^{3N_{\mathrm{el}}}) = \left\{\Psi \in L^2(\mathbb{R}^{3N_{\mathrm{el}}}),\ \|\nabla\Psi\|_{L^2} < +\infty,\ \Psi \text{ antisymmetric}\right\}.$$

As one can see, this space is very high-dimensional, and a naive discretization with 10 points per dimensions leads to $10^{3N_{\mathrm{el}}}$ degrees of freedom, which becomes intractable already when N_{el} exceeds 3 or 4.

In practice, simplified versions of the Schrödinger equation are often employed. One of the most widely used methods today is Density Functional Theory (DFT) [94], which is based on the approximation of the electronic density $\rho_{\mathbf{R}}$, defined as the marginal of the wavefunction

$$\rho_{\mathbf{R}} : \mathbf{r} \in \mathbb{R}^3 \mapsto \rho_{\mathbf{R}}(\mathbf{r}) := N_{\mathrm{el}} \int_{\mathbb{R}^{3(N_{\mathrm{el}}-1)}} |\Psi_{\mathbf{R}}(\mathbf{r}, \mathbf{r}_2, \mathbf{r}_3, \ldots, \mathbf{r}_{N_{\mathrm{el}}})|^2 d\mathbf{r}_2 d\mathbf{r}_3 \ldots d\mathbf{r}_{N_{\mathrm{el}}}. \tag{1.1.3}$$

Physically, it indicates where the electrons are most likely to be present.

Among the various DFT models, the Kohn–Sham model [94] is one of the most used in practice, as it provides an excellent compromise between accuracy and computational cost. This model trades the high-dimensionality of the Schrödinger equation against non-linearity: one needs to compute N_{el} eigenfunctions of a nonlinear operator instead of one for the Schrödinger equation, but those eigenfunctions depend on three variables instead of $3N_{\text{el}}$.

Mathematically speaking, it is very similar to the Hartree–Fock model [76] which consists of restricting the minimization problem (1.1.2) to a small set of wavefunctions called Slater determinants. The Kohn–Sham problem writes: find N_{el} orthonormal eigenfunctions in the Sobolev space $H^1(\mathbb{R}^3)$

$$\Phi_{\mathbf{R}} = (\phi_{\mathbf{R},1}, \dots, \phi_{\mathbf{R},N_{\text{el}}}) \in \left\{ \Phi = (\phi_1, \dots, \phi_{N_{\text{el}}}) \in \left[H^1(\mathbb{R}^3)\right]^{N_{\text{el}}} \middle| \int_{\mathbb{R}^3} \phi_i \phi_j = \delta_{ij} \right\},$$

with corresponding eigenvalues $\lambda_{\mathbf{R},1}, \dots, \lambda_{\mathbf{R},N_{\text{el}}} \in \mathbb{R}$, such that

$$\left(-\frac{1}{2}\Delta + V_{\mathbf{R},\rho_{[\Phi_{\mathbf{R}}]}}\right) \phi_{\mathbf{R},i} = \lambda_{\mathbf{R},i} \phi_{\mathbf{R},i}, \quad i = 1, \dots, N_{\text{el}}, \quad \text{with} \tag{1.1.4}$$

$$\rho_{[\Phi_{\mathbf{R}}]} = \sum_{i=1}^{N_{\text{el}}} |\phi_{\mathbf{R},i}|^2, \tag{1.1.5}$$

where the potential $V_{\mathbf{R},\rho_{[\Phi_{\mathbf{R}}]}}$ depends on the electronic density, which makes the problem non-linear, and is defined as

$$V_{\mathbf{R},\rho} = V_{\mathbf{R}}^{\text{ne}} + V_{\text{coul},\mathbf{R}}(\rho) + V_{\text{xc},\mathbf{R}}(\rho), \tag{1.1.6}$$

where $V_{\mathbf{R}}^{\text{ne}}$ models the interaction between the nuclei and the electrons, $V_{\text{coul},\mathbf{R}}(\rho)$ contains the Coulomb interaction between the electrons, and $V_{\text{xc},\mathbf{R}}(\rho)$ is an additional term called exchange-correlation that models what is missed by the other terms due to the drastic dimensionality reduction.

An important quantity of interest appearing in the numerical resolution described below is the density matrix, which is defined as

$$\begin{aligned} \gamma_{\mathbf{R}} : \ (\mathbb{R}^3)^2 \quad &\to \mathbb{R} \\ (\mathbf{r}, \mathbf{r}') \mapsto \gamma_{\mathbf{R}}(\mathbf{r}, \mathbf{r}') &:= \int_{\mathbb{R}^{3(N_{\text{el}}-1)}} \overline{\Psi_{\mathbf{R}}}(\mathbf{r}, \mathbf{r}_2, \mathbf{r}_3, \dots, \mathbf{r}_{N_{\text{el}}}) \Psi_{\mathbf{R}}(\mathbf{r}', \mathbf{r}_2, \mathbf{r}_3, \dots, \mathbf{r}_{N_{\text{el}}}) d\mathbf{r}_2 d\mathbf{r}_3 \dots d\mathbf{r}_{N_{\text{el}}} \end{aligned} \tag{1.1.7}$$

for the Schrödinger case, or in the Kohn–Sham context as

$$\gamma_{\mathbf{R}} : (\mathbf{r}, \mathbf{r}') \in (\mathbb{R}^3)^2 \mapsto \gamma_{\mathbf{R}}(\mathbf{r}, \mathbf{r}') := \sum_{i=1}^{N_{\text{el}}} \phi_{\mathbf{R},i}(\mathbf{r}) \phi_{\mathbf{R},i}(\mathbf{r}'), \tag{1.1.8}$$

which can also be written as an operator (using the Dirac bra-ket notation) as

$$\gamma_{\mathbf{R}} = \sum_{i=1}^{N_{\text{el}}} |\phi_{\mathbf{R},i}\rangle\langle\phi_{\mathbf{R},i}|.$$

We will also be interested in the pair density defined as

$$\rho_{2,\mathbf{R}} : (\mathbf{r}, \mathbf{r}') \in (\mathbb{R}^3)^2 \mapsto \rho_{2,\mathbf{R}}(\mathbf{r}, \mathbf{r}') := \binom{N_{\text{el}}}{2} \int_{\mathbb{R}^{3(N_{\text{el}}-2)}} |\Psi_{\mathbf{R}}(\mathbf{r}, \mathbf{r}', \mathbf{r}_3, \dots, \mathbf{r}_{N_{\text{el}}})|^2 \quad d\mathbf{r}_3 \dots d\mathbf{r}_{N_{\text{el}}}. \tag{1.1.9}$$

For materials systems, there is an additional difficulty: these systems are modeled as infinite periodic systems. The nuclei are placed on a periodic lattice with unit cell Ω. In practice, the

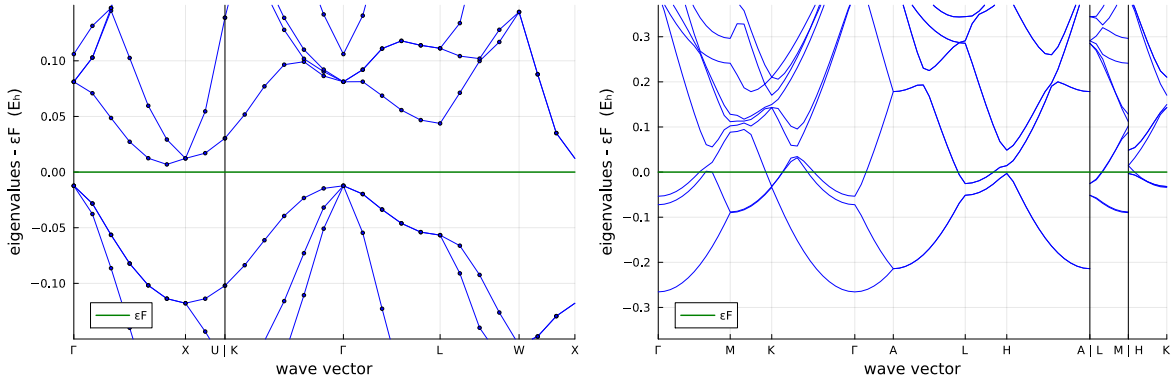


Figure 1.1: Band structure of Silicon (left), which is an insulator, and Magnesium (right), which is a metal. The plots are generated with DFTK.jl [79]. The Fermi level defined in Section 1.1.2 is denoted by ε_F. The wave vectors are $\mathbf{k}$-points.

system is modeled as a so-called thermodynamic limit. We consider larger and larger domains with L^3 copies of the unit cell Ω (L per dimension) called a supercell glued with periodic boundary conditions, with a number of electrons proportional to L^3, and then take L to infinity [69].

For a given L, on the supercell domain, the periodicity of the cell Ω has a specific symmetry which allows to block diagonalize the Hamiltonian and solve smaller problems posed on the unit cell Ω with the modified Hamiltonian

$$H_{\mathbf{R},\mathbf{k}} = \frac{1}{2}(-i\nabla + \mathbf{k})^2 + V_{\mathbf{R},\rho_{[\Phi_{\mathbf{R}}]}},$$

that depends on a parameter $\mathbf{k}$ called $\mathbf{k}$-point, which at the limit $L \to +\infty$ belongs to the Brillouin zone $\Gamma \subset \mathbb{R}^3$ which is the Voronoi cell around the origin of the reciprocal lattice of the material. We denote the eigenvalues for this problem $\lambda_{i,\mathbf{k}}$.

In practice, the Brillouin zone is discretized with a few points per dimension and quadrature rules are used to compute properties integrated over the Brillouin zone Γ. The eigenvalues computed at each $\mathbf{k}$-point form what is called the band structure of the system. They are plotted on Figure 1.1 for two different systems, Silicon and Magnesium, for specific paths in 3D among all possible $\mathbf{k}$-points.

1.1.2 Quantities of interest

In practice, many quantities of interest derive from the density matrix, the energy, or the electronic density. A first simple example is the case of interatomic forces, which are defined as $-\nabla_{\mathbf{R}} E_{\mathbf{R}}$, where $E_{\mathbf{R}}$ is the total energy of the system parametrized by the nuclei positions defined in (1.1.1). In the simulation of materials, one can compute e.g. the integrated density of states, which is defined as the following integral over the Brillouin zone

$$\mathcal{N}(\lambda) = \sum_i \int_\Gamma \mathbb{1}_{\{\lambda_{i,\mathbf{k}} \leq \lambda\}} d\mathbf{k}.$$

The Fermi level, which corresponds to the green lines on the two plots is defined as λ such that $\mathcal{N}(\lambda) = N_{\text{el}}$, where N_{el} is the number of electrons in the cell.

Further, one can also compute more involved properties such as the lattice constant, elastic properties, melting temperatures, or photoemission spectra. For computing some of these

properties, the Kohn–Sham equations typically have to be solved a large number of times for different positions $\mathbf{R}$ of the nuclei. This is the case in molecular dynamics (MD) simulations, where the equations are solved at each step of the time-stepping scheme, the positions of the nuclei being changed according to the forces applied to them, or in a geometry optimization context, where one wants to compute the most stable nuclei configuration of the given molecular system, i.e. the one with the lowest energy.

Electronic structure models are also used a lot to generate precise databases for data-driven models that completely avoid the resolution of the partial differential equations, including e.g. interatomic potentials [21, 12], presented in the next subsection.

1.1.3 Interatomic potentials

For very large systems, electronic structure calculations are often intractable in practice, or at least not for a large number of atomic configurations $\mathbf{R}$. In molecular dynamics calculations, only the total energy of the system as well as the forces applied to the nuclei are needed to run the simulations. Therefore, many methods consist of approximating directly the energy $E_{\mathbf{R}}$ as a function of the nuclei positions and to compute the forces as $-\nabla_{\mathbf{R}} E_{\mathbf{R}}$. There exist many formulations, starting from very simple models, such as the Lennard–Jones potential, which writes

$$E_{\mathbf{R}} = \sum_{i \neq j} f(r_{ij}),$$

with

$$f(r) = \alpha \left[\left(\frac{d}{r} \right)^{12} - \left(\frac{d}{r} \right)^{6} \right],$$

where r_{ij} is the distance between the atoms i and j, and α and d are tabulated parameters of the Lennard–Jones model [100]. This model was originally proposed to describe the interaction between molecules of noble gases, noting that the energy should explode when particles get close, and using a particular exponent to match the London dispersion at long range. The exponent used at short range was fitted to experimental data. Many other interatomic potentials exist, and are based on a functional with few parameters, often fitted to physical quantities.

Since 2007, and the seminal paper [21], many interatomic potentials have been developed based on machine-learning approaches, called Machine-Learned Interatomic Potentials (MLIPs), see e.g. [32, 13, 136, 146, 15]. These interatomic potentials are based on three main ingredients: (i) An accurate enough database, often consisting of electronic structure calculations, where for given structures of atoms, one has computed the corresponding energy and forces, (ii) a set of atomic descriptors which captures the neighborhood of each atom as a vector and (iii) a functional form for the energy that takes as input the descriptors and depends on parameters that are fitted to match the energy and forces in the database.

The choice of the functional form has a large impact on the success of the resulting interatomic potential on at least four aspects. First, the functional form impacts the training of the MLIP, that is the optimization of the free parameters in the energy functional. Indeed between a linear model that can be optimized by solving a least square system, and a deep neural network, the difficulty of solving the optimization problem varies a lot. Second, the final computational cost of evaluating the energy and forces for a given system is key to perform long molecular dynamics simulations. Third, extrapolation capabilities are very important; they are more difficult to obtain for MLIPs than simple interatomic potentials such as the Lennard–Jones, because the number of parameters in the model for the energy is way higher for MLIPs. Four, the exact energy functional $E_{\mathbf{R}}$ from (1.1.1) satisfies symmetry properties. Indeed it is invariant with respect to rigid-body motion, that is translations and rotations of

the inputs, as well as permutations of identical atoms. Therefore these symmetry properties are often imposed on the energy functional form, since forgetting these symmetries may have a large impact on the results (see [26]).

1.2 Discretization and resolution of electronic structure problems

In this section, we first recall the problem of finding the ground state of a molecular or materials system, and detail its numerical resolution, in order to introduce the notation necessary for the following chapters.

For simplicity, compared to (1.1.4), we drop the dependency in the nuclei positions $\mathbf{R}$, as we only consider here a fixed parameter $\mathbf{R}$. Also, we do not mention the $\mathbf{k}$-point dependency for materials systems. Second, although it is usual in the electronic structure community to denote by H the Hamiltonian of the system, we will denote it by A in this section and in Chapter 2, in order to avoid too much notation with the letter h especially relative to Sobolev norms.

1.2.1 Electronic ground-state problem

We focus on the Kohn–Sham DFT ground state problem. Mathematically speaking, it is a minimization problem that can be formulated in the density matrix framework as follows

$$\min\{E(\gamma),\ \gamma \in \mathcal{M}\}, \tag{1.2.1}$$

where

$$\mathcal{M} \coloneqq \{\gamma \in \mathcal{H} \mid \gamma^* = \gamma,\ \gamma^2 = \gamma,\ \mathrm{Tr}(\gamma) = N_{\mathrm{el}},\ \ \mathrm{Ran}\,\gamma \subset \mathcal{V}\} \tag{1.2.2}$$

is the manifold of rank-N_{el} orthogonal projectors (density matrices), $\mathcal{H}$ is a space of operators on some given Hilbert space (to be made more precise later), and $E : \mathcal{H} \to \mathbb{R}$ is a C^2 nonlinear energy functional. The space $\mathcal{V}$ is also a Hilbert space. The parameter N_{el} is a fixed integer depending on the physical model, typically the number of electrons or electron pairs. The unknown is the density matrix, an orthogonal projector of rank N_{el}. The energy functional E is of the form

$$E(\gamma) \coloneqq \mathrm{Tr}(A\gamma) + E_{\mathrm{nl}}(\gamma),$$

where A is the linear part of the mean-field Hamiltonian, and E_{nl} a nonlinear contribution which depends on the considered model. Writing the first-order optimality conditions of (1.2.1), we obtain the following eigenvalue equations, similarly to (1.1.4): Find $\Phi = (\phi_1, \dots, \phi_{N_{\mathrm{el}}}) \in \mathcal{V}$, $\Lambda = (\lambda_1, \dots, \lambda_{N_{\mathrm{el}}}) \in \mathbb{R}^{N_{\mathrm{el}}}$ satisfying

$$\left(A + V_\Phi^{\mathrm{nl}}\right)\phi_i = \lambda_i \phi_i, \quad i = 1, \dots, N_{\mathrm{el}}, \tag{1.2.3}$$

$$\langle \phi_i, \phi_j \rangle = \delta_{ij}, \quad i, j = 1, \dots, N_{\mathrm{el}}, \tag{1.2.4}$$

where V_Φ^{nl} is the nonlinear part of the Hamiltonian arising from $E_{\mathrm{nl}}(\gamma)$, and $\lambda_1, \dots, \lambda_{N_{\mathrm{el}}}$ are the lowest eigenvalues of $A + V_\Phi^{\mathrm{nl}}$, assuming that the Aufbau principle holds [36]. The density matrix γ is linked to the orbitals Φ as follows

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

1.2.2 Discretization of the ground-state problem

The numerical approximation of the solutions to problem (1.2.3) is done by first discretizing the equations on a finite-dimensional space, by means of a Galerkin method. In the conforming approach that we consider in this manuscript, this consists of choosing a finite-dimensional space $\mathcal{V}_N \subset \mathcal{V}$ and solving the following minimization problem:

$$\min\{E(\gamma),\ \gamma \in \mathcal{M}_N\}, \tag{1.2.5}$$

where

$$\mathcal{M}_N \coloneqq \{\gamma \in \mathcal{H} \ \big|\ \gamma^* = \gamma,\ \gamma^2 = \gamma,\ \mathrm{Tr}(\gamma) = N_{\mathrm{el}},\ \ \mathrm{Ran}\,\gamma \subset \mathcal{V}_N\}. \tag{1.2.6}$$

Using the orbital formalism, this amounts to solving the finite-dimensional nonlinear problem: Find $\Phi_N = (\phi_{1,N}, \dots, \phi_{N_{\mathrm{el}},N}) \in \mathcal{V}_N$, $\Lambda_N = (\lambda_{1,N}, \dots, \lambda_{N_{\mathrm{el}},N}) \in \mathbb{R}^{N_{\mathrm{el}}}$ satisfying

$$\begin{gathered}\Pi_N \left(A + V_{\Phi_N}^{\mathrm{nl}}\right) \phi_{i,N} = \lambda_{i,N}\phi_{i,N}, \quad i = 1, \dots, N_{\mathrm{el}}, \\ \langle \phi_{i,N}, \phi_{j,N} \rangle = \delta_{ij}, \quad i,j = 1, \dots, N_{\mathrm{el}},\end{gathered} \tag{1.2.7}$$

where Π_N is the orthogonal projection on $\mathcal{V}_N$ for the scalar product $\mathcal{H}$ is endowed with.

Two typical types of discretization bases are considered: localized bases for molecular systems, and plane waves for materials system. For a textbook on discretization bases, we refer to [116].

For molecular systems, localized basis sets in particular Linear Combinations of Atomic Orbitals (LCAO) are usually preferred. This means that the basis set $\{\chi_\mu\}_{1 \le \mu \le N}$ is composed of functions centered at atomic positions

$$\{\chi_\mu\} = \{\xi_{1,1}(\cdot - \mathbf{R}_1), \dots, \xi_{1,n_1}(\cdot - \mathbf{R}_1); \dots; \xi_{M,1}(\cdot - \mathbf{R}_M), \dots, \xi_{M,n_M}(\cdot - \mathbf{R}_M)\},$$

with $\xi_{i,j} \in \mathrm{H}^2(\mathbb{R}^3)$ called atomic orbitals, $n_1, \dots, n_K \in \mathbb{N}$ such that $\sum_{m=1}^M n_m = N$, the total number of basis functions. Typically, in dimension 3, $\xi_{i,j}(x) = \rho(|x|)Y_l^m(\frac{x}{|x|})$, where ρ is a Gaussian function and Y_l^m is a spherical harmonics (see [77, Chapter 6] for further details). The scalar product in (1.2.4) is then defined on $\mathbb{R}^3$, and one of the main advantages of such basis set is that many quantities appearing in the discretization, in particular the integrals needed to compute the discrete Hamiltonian matrix can be computed analytically. A major drawback is that in practice it is very difficult to converge numerical results to machine precision because the basis sets very quickly become ill-conditioned. Also it is in general not known how to systematically and optimally choose the exponents of the Gaussian functions as well as the degrees of the spherical harmonics to reach a given accuracy.

In the periodic setting used e.g. for materials systems, plane waves are often used. Let us detail a bit the functional setting in this case. Let $\Omega \subset \mathbb{R}^3$ denote a unit cell of an arbitrary periodic lattice $\mathcal{R}$, with reciprocal lattice $\mathcal{R}^*$. The space of complex-valued, 2-integrable $\mathcal{R}$-periodic functions

$$L^2_{\mathrm{per}}(\Omega) = \{u \in L^2_{\mathrm{loc}}(\mathbb{R}^3; \mathbb{C}) : u \text{ is } \mathcal{R}\text{-periodic}\}, \tag{1.2.8}$$

admits an orthonormal basis consisting of plane waves:

$$e_{\boldsymbol{G}} : \boldsymbol{x} \in \mathbb{R}^3 \mapsto |\Omega|^{-\frac{1}{2}} e^{i\boldsymbol{G}\cdot\boldsymbol{x}},\ \boldsymbol{G} \in \mathcal{R}^*. \tag{1.2.9}$$

The discretization space is defined, given a discretization parameter $E_{\mathrm{cut}} \in \mathbb{N}$, for $N_{\mathrm{cut}} = \sqrt{2E_{\mathrm{cut}}}$, by

$$\mathcal{V}_{N_{\mathrm{cut}}} = \mathrm{Span}\left(e_{\boldsymbol{G}} :\ |\boldsymbol{G}| \le N_{\mathrm{cut}}\right) = \mathrm{Span}\left(e_{\boldsymbol{G}} :\ \frac{1}{2}|\boldsymbol{G}|^2 \le E_{\mathrm{cut}}\right). \tag{1.2.10}$$

The parameter E_{cut} appearing in the previous definition is known in the materials science community as the energy cutoff. The main advantage of plane waves is that the Laplace operator is diagonal in this basis and fast Fourier transforms can be used to efficiently compute the discretization matrix of the Hamiltonian, in particular the potential part of the operator.

In many other fields finite elements are used to discretize partial differential equations. But they are seldom used in electronic structure calculations. As a main advantage, the operator and mass matrices are sparse, but a large number of basis functions is needed to obtain accurate results. In some cases, adaptive finite elements mitigate this.

1.2.3 Numerical resolution of the discretized equations

Once a discretization space $\mathcal{V}_N$ has been chosen, equations (1.2.7) need to be solved in practice. If $V_{\Phi_N}^{\text{nl}} = 0$ then the problem is a linear matrix eigenvalue problem for which there exist standard resolution methods. Since the eigenvalue problem is large in general, iterative methods such as the Arnoldi, the Lanczos method or Locally Optimal Block Preconditioned Conjugate Gradient (LOBPCG) are prefered [93, 129].

If $V_{\Phi_N}^{\text{nl}} \neq 0$ the problem is a nonlinear matrix eigenvalue problem, and is usually solved by means of an iterative algorithm such as the self-consistent field (SCF) algorithm [127]. Indeed once a finite discretization basis of size N has been chosen, the nonlinear eigenvalue matrix problem reads: Find eigenvectors $C \in \mathbb{R}^{N \times N_{\text{el}}}$ and eigenvalues $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_{N_{\text{el}}})$ such that

$$\begin{aligned} F(CC^T)C &= SC\Lambda \\ C^T SC &= I_{N_{\text{el}}}, \end{aligned} \tag{1.2.11}$$

where the matrix

$$D = CC^T, \tag{1.2.12}$$

which is the S-orthogonal projector onto the N_{el} orbitals, is the discrete version of the density matrix (1.1.7). The so-called Fock matrix $F(CC^T)$ contains the discretization of the Hamiltonian $A + V_{\Phi_N}^{\text{nl}}$, S is the overlap matrix of the basis, $I_{N_{\text{el}}}$ denotes the identity matrix of size N_{el}. The SCF algorithm consists in starting from an initial guess $C^{(0)}$ and solving the following linear eigenvalue problem at each iteration k

$$\begin{aligned} F(C^{(k-1)}{C^{(k-1)}}^T)C^{(k)} &= S^{(k)}C^{(k)}\Lambda^{(k)} \\ (C^{(k)})^T S^{(k)} C^{(k)} &= I_{N_{\text{el}}}, \end{aligned} \tag{1.2.13}$$

where the non-linearity is frozen at the previous step. The iterations are stopped when a given convergence criterium is met. To ensure and often accelerate convergence, a mixing step is typically included. Without it, the algorithm may fail to converge. After $k_{\max}$ iterations, the approximate solution in the discretization basis $\{\chi_n\}_{n=1,\ldots,N}$ is

$$\Phi_{N,k_{\max}} = (\phi_{1,N,k_{\max}}, \ldots, \phi_{N_{\text{el}},N,k_{\max}}),$$

with

$$\forall i = 1, \ldots N_{\text{el}}, \quad \phi_{i,N,k_{\max}} = \sum_{n=1}^{N} C_{ni}\chi_n.$$

The corresponding density matrix is

$$\gamma_{N,k_{\max}} = \sum_{i=1}^{N_{\text{el}}} |\phi_{i,N,k_{\max}}\rangle\langle\phi_{i,N,k_{\max}}|. \tag{1.2.14}$$

Direct minimization algorithms also exist and consist in solving directly the nonlinear discrete minimization problem (1.2.5). A comparison between SCF and direct minimization is presented in [38].

Due to the diagonalizations of the Hamiltonian matrix, the algorithms typically scale cubically with respect to the number of basis functions N. The overall cost then depends mainly on three elements: (i) the number of basis functions used (ii) the cost of assembling the Fock matrix at each iteration (iii) the number of iterations needed to reach convergence.

1.3 Research context summary

To summarize the context so far, in electronic structure calculations, we face high-dimensional (possibly nonlinear) eigenvalue equations that are parametrized by the nuclei positions and that have to be solved a large number of times for different positions of the nuclei. These equations are numerically very costly to solve, and several approximations have to be considered in the resolution of the problems. Then different techniques can be resorted to for approximating the solutions efficiently among which developing error bounds, reduced order models, and machine learning techniques.

Error estimation When we solve an electronic structure problem, we start with choosing a model, that is then discretized and practically solved using iterative algorithms. This raises a number of important questions. (1) Once a numerically computed solution is obtained, what is the error between the exact solution of the considered equations and the computed solution? Answering this question is key to know whether the computed solution is accurate enough for applications. *A posteriori* error estimation is a great tool to provide such error bounds, that are, moreover, guaranteed and computable. (2) How much does the error depend on each parameter (e.g. discretization) of the simulation? Answering this question allows to adaptively refine the parameters of the simulations, and decrease the overall computational cost of computing a solution with a prescribed accuracy.

Reduced order modeling Controlling the error between the exact and approximate solutions for single point calculations can be used to accelerate the simulations for a given configuration parameter $\mathbf{R}$. However one often needs to efficiently approximate a whole manifold of solutions. In our context this can be electronic densities

$$\mathcal{M}^{\rho} := \{\rho_{\mathbf{R}}, \quad \mathbf{R} \in \mathcal{C}\}$$

or density matrices

$$\mathcal{M}^{\gamma} := \{\gamma_{\mathbf{R}}, \quad \mathbf{R} \in \mathcal{C}\},$$

where $\mathcal{C}$ is a meaningful set of configurations that will depend on the context. In this direction, the questions are mainly concerned with the efficient approximation of the manifold, from the knowledge of a few points in the manifold. This is closely related to the development of reduced order models. As will be described later on, linear reduced order models are not well suited for this application, so nonlinear reduced order models are resorted to. This raises the question of how to develop accurate and reliable reduced order models in this context.

Machine learning When it comes to data-driven approaches, the questions considered in this manuscript are slightly different. We are looking for adapted parametrizations of high-dimensional functionals, typically the potential energy $E_{\mathbf{R}}$ so that the computational cost

of evaluating the functional $E_{\mathbf{R}}$ is low, and that the training of the functional can be done reasonably easily. We are also interested in how these techniques can be further adapted to electronic structure models, and in particular how such parametrizations can be adapted to electronic density, Hamiltonians, or wavefunction learning. My interest in this field therefore goes toward the development of efficient descriptors of atomic environments, and studying approximation theory under symmetry constraints.

1.4 Contributions

In this section I briefly describe my contributions and present a list of publications.

1.4.1 Summary of the contributions

In Chapters 2 to 5, I summarize the contributions made after the PhD thesis. Note that some of the contributions, as detailed below, are not mentioned in the rest of the manuscript.

In Chapter 2, I describe the contributions to the field of *a posteriori* error estimation for eigenvalue problems and the development of efficient numerical methods. The main idea is to estimate the error between the exact and computed solution of a partial differential equation that is in this context an eigenvalue problem. Several difficulties are in order, such as the non-linearity of the operator, if it depends of the exact eigenvectors as is the case in DFT, the computation of clusters of eigenvalues, or the possibly non-self-adjointness of the operator. First, a review on *a posteriori* error analysis and post-processing methods for nonlinear eigenvalue problems has been published in [A20]. Regarding the computation of clusters of eigenvalues, fully guaranteed and computable error bounds are presented in [A10], a work that has been started during my PhD but finished afterwards. The discretization methods considered there are finite elements or plane waves. A fully guaranteed and computable error estimation for nonlinear eigenvalue problems with convex density functional such as the reduced Hartree–Fock model is presented in [A3]. The numerical results are presented in the case of a plane wave discretization. Recently, a first guaranteed error estimation together with adaptive basis sets has been provided in the case of a discretization based on localized basis sets [A16]. An optimization strategy for localized basis sets is, moreover, presented in [P1] (omitted later). Error bounds for quantities of interest, such as interatomic forces are presented in [A5].

On a different topic linked to the simulation of a nuclear core, we have considered non-self-adjoint operators and provided error bounds that can be used in a reduced basis context [A12] (omitted later), reducing the overall computation cost of the simulations.

I also present contributions aiming at rendering numerical methods in quantum chemistry more efficient, but that do not directly rely on fully guaranteed error bounds. A work on adaptive cutoff choice in plane wave methods in electronic structure calculations is presented in [A23] (omitted later). Several works are based on perturbation theory, such as two articles discussing the Feschbach–Schur map [A22, A21], and a multipoint perturbation theory [A17].

My co-authors for these works are Andrea Bordignon, Eric Cancès, Huajie Chen, Yonah Conjungo Taumhas, Mi-Song Dupuy, Virginie Ehrlacher, Jun Fang, Xingyu Gao, Louis Garrigue, Gaspard Kemlin, Antoine Levitt, Tony Lelièvre, Filippo Lipparini, Beilei Liu, Ioanna-Maria Lygatsika, Yvon Maday, François Madiot, Rafael Antonio Lainez Reyes, Michael Sigal, Benjamin Stamm, Laurent Vidal, Martin Vohralík.

In Chapter 3 are presented several contributions related to optimal transport, that aim at proposing distances adapted to electronic structure calculations. We present the development of modified Wasserstein barycenters for mixtures of probability distributions in [S2], as well as modified Wasserstein barycenters that satisfy given marginal properties [A13].

My co-authors for these works are Maxime Dalery, Virginie Ehrlacher, and Nathalie Nouaime.

In Chapter 4 are presented contributions to nonlinear model order reduction for quantum chemistry-based problems. A numerical method aiming at reducing the computational cost of *ab initio* molecular dynamics simulations based on an extrapolation scheme on the Grassmann manifold was presented in [A26, A25, A24]. It substantially reduces the computational time of these simulations. We also propose nonlinear reduced order models based on optimal transport. A first work on a toy model is presented in [S1]. Finally an original view of pair-density approximation based on a statistical object called copula, which is also linked to optimal transport, is presented in [A18].

My co-authors for these works are Maxime Dalery, Virginie Ehrlacher, Gero Friesecke, Claudia Klüppelberg, Filippo Lipparini, Alexei Lozinski, Patrizia Mazzeo, Aleksandr Mikhalev, Federica Pes, Etienne Polack, Benjamin Stamm.

The last chapter of this manuscript is concerned with data-driven methods for the computation of interatomic potentials, as well as Hamiltonian operators, and wavefunctions. Regarding the interatomic potentials, contrarily to the previous chapters, the main interest is in approximating directly the energy and the forces of the system without the use of the electronic structure. The main contributions are based on polynomial approximations of the potential energy surface. A first approximation based on permutation-invariant polynomials and invariant theory is presented in [A1, A27]. A more efficient approximation, still based on invariant polynomials and called Atomic Cluster Expansion (ACE) is presented in [A15]. The ACE method has then been extended to the learning of Hamiltonians in [A28]. Theoretical results on the approximation of a class of functions that satisfy permutation invariance, that includes multiset functions can be found in [A2]. To conclude, a recent work [P2] compares different state-of-the-art methods to compute the electronic structure in a data-driven way based on machine-learning and tensor methods.

My co-authors for these works are Alice Allen, Gautam Anand, Markus Bachmayr, Gábor Csányi, Mathias Dus, Clément Guillot, James Kermode, Adam McSloy, Berk Onat, Cas van der Oord, Christoph Ortner, Reinhard Maurer, Joel Pascal Soffo Wambo, Jack Thomas, Liwei Zhang.

1.4.2 List of publications

The articles [A14, A11, A9, A8, A4, A19, A7, A6] were done during my PhD thesis and will not be presented in the following chapters.

Submitted preprint articles

[S1] M. Dalery, G. Dusson, V. Ehrlacher, and A. Lozinski, *Nonlinear reduced basis using mixture Wasserstein barycenters: application to an eigenvalue problem inspired from quantum chemistry*, arXiv:2307.15423, (2023).

[S2] G. Dusson, V. Ehrlacher, and N. Nouaime, *A Wasserstein-type metric for generic mixture models, including location-scatter and group invariant measures*, arXiv:2301.07963, (2023).

Journal articles

[A1] A. E. A. Allen, G. Dusson, C. Ortner, and G. Csányi, *Atomic permutationally invariant polynomials for fitting molecular force fields*, Machine Learning: Science and Technology, 2, 025017 (2021).

[A2] M. Bachmayr, G. Dusson, C. Ortner, and J. Thomas, *Polynomial approximation of symmetric functions*, Mathematics of Computation, 93 (2024), pp. 811–839.

[A3] A. Bordignon, G. Dusson, E. Cancès, G. Kemlin, R. A. L. Reyes, and B. Stamm, *Fully guaranteed and computable error bounds on the energy for periodic kohn-sham equations with convex density functionals*, arXiv 2409.11769, accepted in SIAM Journal of Scientific Computing, (2025).

[A4] E. Cancès and G. Dusson, *Discretization error cancellation in electronic structure calculation: toward a quantitative study*, ESAIM: Mathematical Modelling and Numerical Analysis, 51 (2017), pp. 1617–1636.

[A5] E. Cancès, G. Dusson, G. Kemlin, and A. Levitt, *Practical error bounds for properties in plane-wave electronic structure calculations*, SIAM Journal of Scientific Computing, 44 (2022), pp. B1312–B1340.

[A6] E. Cancès, G. Dusson, Y. Maday, B. Stamm, and M. Vohralík, *A perturbation-method-based a posteriori estimator for the planewave discretization of nonlinear Schrödinger equations*, Comptes Rendus Mathematique, 352 (2014), pp. 941–946.

[A7] E. Cancès, G. Dusson, Y. Maday, B. Stamm, and M. Vohralík, *A perturbation-method-based post-processing for the planewave discretization of Kohn–Sham models*, Journal of Computational Physics, 307 (2016), pp. 446–459.

[A8] E. Cancès, G. Dusson, Y. Maday, B. Stamm, and M. Vohralík, *Guaranteed and robust a posteriori bounds for Laplace eigenvalues and eigenvectors: conforming approximations*, SIAM Journal on Numerical Analysis, 55 (2017), pp. 2228–2254.

[A9] E. Cancès, G. Dusson, Y. Maday, B. Stamm, and M. Vohralík, *Guaranteed and robust a posteriori bounds for Laplace eigenvalues and eigenvectors: a unified framework*, Numerische Mathematik, 140 (2018), pp. 1033–1079.

[A10] E. Cancès, G. Dusson, Y. Maday, B. Stamm, and M. Vohralík, *Guaranteed a posteriori bounds for eigenvalues and eigenvectors: multiplicities and clusters*, Mathematics of Computation, 89 (2020), pp. 2563–2611.

[A11] E. Cancès, G. Dusson, Y. Maday, B. Stamm, and M. Vohralík, *Post-processing of the planewave approximation of Schrödinger equations. Part I: linear operators*, IMA Journal of Numerical Analysis, 41 (2020), pp. 2423–2455.

[A12] Y. Conjungo Taumhas, G. Dusson, V. Ehrlacher, T. Lelièvre, and F. Madiot, *Reduced basis method for non-symmetric eigenvalue problems: application to the multigroup neutron diffusion equations*, ESAIM: Mathematical Modelling and Numerical Analysis, 58 (2024), pp. 1959–1987.

[A13] M. Dalery, G. Dusson, and V. Ehrlacher, *Marginal-constrained modified Wasserstein barycenters for Gaussian distributions and Gaussian mixtures*, hal-04696783, accepted in SIAM Journal of Matrix Analysis, (2025).

[A14] G. Dusson, *Post-processing of the plane-wave approximation of Schrödinger equations. Part II: Kohn–Sham models*, IMA Journal of Numerical Analysis, 41 (2020), pp. 2456–2487.

[A15] G. Dusson, M. Bachmayr, G. Csányi, R. Drautz, S. Etter, C. van der Oord, and C. Ortner, *Atomic cluster expansion: Completeness, efficiency and stability*, Journal of Computational Physics, 454 (2022), p. 110946.

[A16] G. Dusson, M.-S. Dupuy, and I.-M. Lygatsika, *A posteriori error estimates for Schrödinger operators discretized with linear combinations of atomic orbitals*, arXiv:2410.04943, accepted in SIAM Journal of Numerical Analysis, (2025).

[A17] G. Dusson, L. Garrigue, and B. Stamm, *A multipoint perturbation formula for eigenvalue problems*, ESAIM: Mathematical Modelling and Numerical Analysis, 59 (2025), pp. 2081–2109.

[A18] G. Dusson, C. Klüppelberg, and G. Friesecke, *Copula methods for modeling pair densities in density functional theory*, Journal of Chemical Physics, 162 (2025), p. 144109.

[A19] G. Dusson and Y. Maday, *A posteriori analysis of a nonlinear Gross–Pitaevskii-type eigenvalue problem*, IMA Journal of Numerical Analysis, (2016), p. drw001.

[A20] G. Dusson and Y. Maday, *An overview of a posteriori error estimation and post-processing methods for nonlinear eigenvalue problems*, Journal of Computational Physics, 491 (2023), p. 112352.

[A21] G. Dusson, I. Sigal, and B. Stamm, *Analysis of the Feshbach–Schur method for the fourier spectral discretizations of schrödinger operators*, Mathematics of Computation, 92 (2023), pp. 217–249.

[A22] G. Dusson, I. M. Sigal, and B. Stamm, *The Feshbach–Schur map and perturbation theory*, in the book *Partial Differential Equations, Spectral Theory, and Mathematical Physics: The Ari Laptev Anniversary Volume*, EMS Series of Congress Reports (2021).

[A23] B. Liu, H. Chen, G. Dusson, J. Fang, and X. Gao, *An adaptive planewave method for electronic structure calculations*, SIAM Multiscale Modeling and Simulation, 20 (2022), pp. 524–550.

[A24] F. Pes, É. Polack, P. Mazzeo, G. Dusson, B. Stamm, and F. Lipparini, *A quasi time-reversible scheme based on density matrix extrapolation on the Grassmann manifold for Born-Oppenheimer molecular dynamics*, Journal of Physical Chemistry Letters, 14 (2023), pp. 9720–9726.

[A25] E. Polack, G. Dusson, B. Stamm, and F. Lipparini, *Grassmann extrapolation of density matrices for Born–Oppenheimer molecular dynamics*, Journal of Chemical Theory and Computation, 17 (2021), pp. 6965–6973.

[A26] É. Polack, A. Mikhalev, G. Dusson, B. Stamm, and F. Lipparini, *An approximation strategy to compute accurate initial density matrices for repeated self-consistent field calculations at different geometries*, Molecular Physics, 118 (2020), p. e1779834.

[A27] C. van der Oord, G. Dusson, G. Csányi, and C. Ortner, *Regularised atomic body-ordered permutation-invariant polynomials for the construction of interatomic potentials*, Machine Learning: Science Technology, 1 (2020), p. 015004.

[A28] L. Zhang, B. Onat, G. Dusson, A. McSloy, et al., *Equivariant analytical mapping of first principles hamiltonians to accurate and transferable materials models*, npj Computational Materials, 8 (2022), p. 158.

Proceedings

[P1] É. Cancès, G. Dusson, G. Kemlin, and L. Vidal, *On basis set optimisation in quantum chemistry*, ESAIM: Proceedings and Surveys, abs/2207.12190 (2022).

[P2] M. Dus, G. Dusson, V. Ehrlacher, C. Guillot, and J. P. S. Wambo, *Comparison between tensor methods and neural networks in electronic structure calculations*, arXiv:2412.03319, accepted in ESAIM: Proceedings and Surveys, (2024).

Chapter 2

A posteriori error estimation and efficient numerical methods for eigenvalue problems

The aim of *a posteriori* error estimation is to provide guaranteed and computable error bounds between the exact solution of a given equation and a numerically computed – hence approximate – solution to this equation. As we will see in this chapter, these *a posteriori* error bounds can in general be used for two purposes, either to guarantee the accuracy of the computed solution, or to improve the computed solution at a low computational cost.

The partial differential equations of interest in this manuscript are (possibly nonlinear) eigenvalue problems. In this chapter, we start by presenting in Section 2.1 the main ideas behind *a posteriori* error estimation, and the available tools to estimate the error between exact and approximate solutions. The following Section 2.2 details different aspects of *a posteriori* error bounds and how the contributions [A10, A20, A5, A16, A3] fit in the literature. In Section 2.3, we present contributions concerning perturbation theory, a great tool for obtaining approximations of solutions to problems that are close to known equations. In Section 2.4, we present research perspectives related to this subject.

2.1 *A posteriori* error estimation for generic problems

To present the gist of *a posteriori* error estimation, we start by considering a generic problem: find $x^* \in V$ such that

$$F(x^*) = 0,$$

where $F : V \to W$ is a smooth nonlinear mapping and V and W are two Banach spaces. In general, only numerical approximations $x \in V$ of a solution x^* are available. Often, the numerical analysis starts by deriving an *a priori* error estimation, which provides a convergence rate of the error between the approximate solutions x and the exact solution x^* that depends on the number of parameters used to represent the solution x, for a given numerical method. However this is in general insufficient to guarantee that the error is below a given tolerance, e.g. 10%, 5%, 1%, because of unknown quantities in the estimations, such as the norm of the exact solution. The goal of *a posteriori* error estimation is therefore to bound the error $\|x - x^*\|$ by quantities that are both guaranteed and computable. In principle, this leads to a control over the approximation error which allows to ensure that indeed the numerical error is below a prescribed tolerance.

The main available tool in this field is the residual analysis. The residual quantifies the

error in satisfying the equation, that is $F(x)$. It is zero for the exact solution x^* but nonzero for a generic x. Naturally we expect that a small norm of the residual $\|F(x)\|$ corresponds to a small error $\|x - x^*\|$. The aim of the *a posteriori* error estimation is to transform this vague statement into a rigorous one, with computable quantities.

To give an example of how this can be done, we consider the toy problem of solving a linear system $Ax^* = b$, with A an n by n real invertible matrix, and b a vector of size n. The residual function F is in this case $F : \mathbb{R}^n \to \mathbb{R}^n$ with

$$F(x) = Ax - b.$$

Then, given an approximate solution x, one can remark that since $F(x^*) = 0$,

$$A(x - x^*) = F(x).$$

Therefore, one can easily relate the solution error $x - x^*$ to the residual $F(x)$ e.g. for the 2-norm as

$$\|A(x - x^*)\|_2 = \|F(x)\|_2,$$

which is already an *a posteriori* error bound.

The choice of the norm in the previous equation is in fact crucial, and depends on the norm of interest in the solution error. For example, if we are interested in the 2-norm of the error $x - x^*$, then one has

$$\|x - x^*\|_2 = \|A^{-1}F(x)\|_2,$$

which seems at first sight an excellent *a posteriori* error bound, since there is an equality between left-hand side and right-hand side, and not just an inequality. However, computing $A^{-1}F(x)$ amounts to solving a linear system which is exactly what is needed to find the exact solution x^*, and therefore this error bound is not computable in practice. At that stage, one possibility is to estimate the norm of the matrix A^{-1} which in some cases is known in advance and to write the *a posteriori* error estimation

$$\|x - x^*\|_2 \leq \|A^{-1}\|\|F(x)\|_2, \tag{2.1.1}$$

but this can lead to a huge overestimation of the error, which would then be useless in practice. In between inverting the whole system and hugely overestimating the error, many possibilities can be imagined, and this is where the core work is being done. Indeed, at the end, a good *a posteriori* error bound is a bound that provides an accurate approximation of the error while being reasonably cheap to evaluate in practice.

The link between the error and the residual was particularly simple in the above case due to the linearity of the considered problem. For a generic nonlinear problem $F(x) = 0$, similar estimations can be obtained with a first-order Taylor expansion of the residual. Namely, since $F(x^*) = 0$, one has, up to higher order terms

$$F(x) \approx F(x^*) + DF_x(x - x^*) = DF_x(x - x^*),$$

where DF_x is the differential of F at x. Therefore, in the case where the previous equation can be inverted the error can be approximated by

$$x - x^* \approx [DF_x]^{-1}F(x). \tag{2.1.2}$$

At that stage where the error is estimated as

$$x - x^* \approx \text{err},$$

there are two possibilities depending on the context. One is to compute an improved solution by setting

$$\widetilde{x} := x - \mathrm{err}. \tag{2.1.3}$$

This is the idea behind many standard iterative algorithms such as the gradient method or Newton algorithm, the latter corresponding to the error estimation (2.1.2). The other possibility is to transform the error approximation into a rigorous error bound. This is exactly the result of the Newton–Kantorovitch theorem [89, 115] (implicit function theorem) presented in the functional setting of interest in [34]. Under the assumptions that the differential of F is invertible at x, does not locally vary too much in a neighborhood of x, and assuming that the residual is small enough, then the error can be *a posteriori* estimated, as stated below.

Theorem 2.1.1 (Theorem 2.1 in [34])**.** *Let V, W be two Banach spaces. Let $F : V \to W$ be a C^1 mapping and $x \in V$ be such that $DF_x \in \mathcal{L}(V, W)$ is an isomorphism. Let*

$$\begin{aligned} \varepsilon &= \|F(x)\|_W, \\ \gamma &= \|[DF_x]^{-1}\|_{W,V}, \\ L(\alpha) &= \sup_{y \in \bar{B}(x,\alpha)} \|DF_y - DF_x\|_{V,W}. \end{aligned}$$

Assume that $2\gamma L(2\gamma\varepsilon) \leq 1$. Then the problem $F(x) = 0$ has a unique solution x^* in the ball $\bar{B}(x, 2\gamma\varepsilon)$ and

$$\|x - x^*\|_V \leq 2\gamma \|F(x)\|_W. \tag{2.1.4}$$

This theorem provides a systematic way to obtain *a posteriori* error bounds for nonlinear problems. It essentially amounts to guarantee that the remainder in the first-order Taylor expansion is small enough and can be absorbed in the factor 2 in (2.1.4). However, this theorem has the same drawback as (2.1.1), i.e. the estimate depends on the norm of the inverse of the differential which can hugely overestimate the error in practice. Therefore, the choice of the functional spaces V, W is crucial to obtain accurate bounds. Note that often, the error to be controlled is given by the context, which can also lead to norms in the residual that are difficult to evaluate, typically dual norms. Hence deriving good *a posteriori* error bounds for nonlinear problems remains a challenge in general, and has to be adapted to the equations of interest.

When the considered function F is the gradient of an energy functional that is minimized, as it is the case for the ground state energy, we have the following error estimate on the energy

$$E(x) - E(x^*) \approx \frac{1}{2}\, D^2E_x(x - x^*, x - x^*),$$

where D^2E_x is the second-order differential of the energy E at x. This relation implies that the error in the energy is second-order with respect to the error in the solution. This property will later be exploited in the context of eigenvalue problems, where the eigenvalues are associated with the energy, and the eigenvectors correspond to the solution x.

As an extension, it is natural to wonder how to obtain error bounds on quantities of interest, starting with the simplest case in electronic structure calculations, which is the interatomic forces (see [A5]). Assume that we are interested in a real-valued quantity of interest $Q(x_*)$, where $Q : \mathbb{R}^n \to \mathbb{R}$ is a C^1 function, then we have the approximate equality with computable right-hand side:

$$Q(x) - Q(x_*) \approx \nabla Q(x) \cdot \left(DF_x^{-1} F(x)\right). \tag{2.1.5}$$

From here, we obtain the simple estimate

$$|Q(x) - Q(x_*)| \lesssim |\nabla Q(x)| \left\| DF_x^{-1} \right\|_{\mathrm{op}} \|F(x)\|,$$

where $\|\cdot\|_{\mathrm{op}}$ is the induced operator norms on $\mathbb{R}^{n\times n}$ (note that $\nabla Q(x) \in \mathbb{R}^n$ and $DF_x \in \mathbb{R}^{n\times n}$). This approximate bound can be turned into a rigorous one provided that one can estimate second-order derivatives using Newton–Kantorovitch theorem; see for instance [132]. However in general we face the same problem as above for the solution, which is the overestimation of the error if standard operator norms are used.

Once an *a posteriori* error bound has been obtained, we are able to guarantee that the approximation error is below a given tolerance, which amounts to certify the results of the simulations to some accuracy. We can also use the error bound to adapt the parameters of the simulations in order to reduce the total computational cost leading to a final given accuracy. This requires an additional ingredient, which is the separation of the *a posteriori* error bound into several components that mainly depend on a specific parameter of the simulation. In finite element simulations, the total error is usually separated in local components as

$$\mathrm{err}_{\mathrm{total}} \leq \sum_k (\eta_k)^2,$$

where η_k are local quantities, so that only parts of the mesh where the associated error is the largest are refined [48, 40, 58]. Among the strategies used to determine which elements to refine is the Dörfler marking strategy [54], where only a given proportion of the elements is refined, linked to a proportion of the total error. Although very common in finite element simulations, adaptive strategies are less developed when it comes to local basis sets, or adaptive refinement via an error separation with respect to the discretization on one side and the number of iterations in the iterative algorithm used to compute the solution on the other side. But there are some recent works in this direction, both in the context of electronic structure [147], or for other problems [33, 107].

To conclude, a wish list for the derived *a posteriori* error bounds is the following: (i) computable, (ii) guaranteed, (iii) cheap to evaluate, and (iv) separable with respect to the parameters of the simulations.

2.2 *A posteriori* error estimation for eigenvalue problems

2.2.1 Overview

In my research, I have been interested in electronic structure problems, which often exhibit nonlinear eigenvalue problems, such as the Kohn–Sham model (1.1.4). In electronic structure calculations, the differential operators involved are in most cases, and at least in what will be mentioned in this discussion, self-adjoint, and of the form $-\Delta + V$, where V is some potential. Deriving *a posteriori* error bounds for eigenvalue problems has been a subject of interest at least from the 1960s. Many error bounds have been proposed since, and I summarize my contributions in the field in Table 2.1. Before diving more precisely into the literature and my contributions, I will first provide an overview of the various problem settings considered, along with the main techniques employed to derive *a posteriori* error estimates. From a methodological perspective, one must distinguish between the following categories: (i) the considered operator, (ii) the number and position in the spectrum of eigenvalues that have to be computed, (iii) the discretization basis, (iv) the algorithm used to solve the discretized equations, and (v) the quantities of interest to be computed (e.g., eigenvalues, eigenvectors, derived quantities such as interatomic forces).

Equation / Discretization	Linear single eigenvalue	Linear cluster of eigenvalues	Nonlinear convex single eigenvalue	Nonlinear convex cluster of eigenvalues	Nonlinear nonconvex (not guaranteed)
Finite Elements	[A8]	[A10]			
Plane waves	[A8]	[A10]	[A19]	[A3]	[A5]
Localized basis sets	[A16]				

Table 2.1: Summary of references on *a posteriori* error estimation for eigenvalue problems

The choice of the differential operator, in particular whether it is linear or nonlinear, has a significant impact on the techniques used to derive *a posteriori* error bounds. For linear eigenvalue problems, we can rely on *a posteriori* error bounds that are specific to this setting, as detailed below. In contrast, nonlinear eigenvalue problems often require tools such as Newton–Kantorovich theorem. Additionally in the nonlinear case, the nature of the energy functional plays a crucial role in the feasibility of obtaining guaranteed error estimates. For example, we have established guaranteed bounds in cases involving convex nonlinear energy functionals. However, for nonconvex functionals, deriving globally guaranteed error bounds remains out of reach. That said, for local solutions, it is still possible to obtain meaningful error estimates by analyzing the energy locally. The main difficulty then consists of estimating the norms of the differential of the residual F and choosing relevant Banach spaces. Typically, for the considered operators, the appropriate norms are the Sobolev H^1 norm for the solution space and the dual space H^{-1} for computing the residual.

Some problems involve computing a single eigenvalue, while others a cluster or degenerate eigenvalues. For the latter, the associated eigenvectors may no longer be uniquely defined (even up to a sign), as any orthonormal basis of the eigenspace is valid. Therefore, instead of focusing on individual eigenvectors, it is more appropriate to consider the orthogonal projector onto the subspace spanned by the corresponding eigenvectors. This changes the functional analysis setting but does not change significantly the techniques used in the proofs. But having a gap between the considered eigenvalues and the rest of the spectrum is key to obtain guaranteed error bounds, as it guarantees the uniqueness of the orthogonal projector. In the case where there is no gap, we do not know how to provide error bounds, apart from computing additional eigenvalues and relying on a further gap in the spectrum. Note that in electronic structure, the ground state problem amounts to find the lowest N_{el} eigenvalues of the Hamiltonian operator, but there are other settings of interest where one is interested in other eigenvalues, such as for the computation of excited states.

The choice of the discretization, as mentioned before, often depends on whether molecules or materials systems are considered. In both cases, the discretization is conforming (that is $\mathcal{V}_N \subset \mathcal{V}$) which means that the variational principle holds: the total energy is minimized on a subspace of the space on which the energy is defined. As a result, we therefore have a trivial guaranteed upper bound on the energy and the eigenvalues, which are the numerically computed quantities. This also helps to avoid spectral pollution and ensures that the presence of a continuous spectrum does not pose a problem as long as the highest computed eigenvalue is below the bottom of the continuous spectrum [A16].

When a plane wave discretization is considered, as is usually done for materials systems, a key observation is that the Laplace operator is diagonal on the discretization basis. This means that all Sobolev norms including the dual norms $H^{-s}, s > 0$ can be computed exactly very easily. Moreover the potential is asympotically small with respect to the Laplace operator,

especially when it is smooth. Therefore we often separate between low frequencies and high frequencies, and neglect the potential outside of the discretization basis. Different flavors of this argument can be found in the described contributions below: one can use perturbation theory where the potential in the high frequencies is seen as a perturbation of the whole operator, or use a Schur complement and only solve a linear system in the low-frequency space.

For the molecular case discretized with a localized basis, these arguments are not valid anymore. However the atomic problems, i.e. $-\Delta + V$ with V centered on a single atom are precisely solved with atomic basis sets from the literature. This can be exploited to obtain guaranteed error bounds, and especially estimate the dual norms appearing in the estimations. Using these bounds, one can construct adaptive basis sets, similarly to finite elements adaptive refinement techniques. The main difficulty is that the bases are not compactly supported so the error has to be made local in some way. For this we used a partition of unity.

Compared to my PhD thesis where I mostly focused on the plane wave discretization error for linear eigenvalue problems, and some nonlinear eigenvalue problem in 1D, I since focused on many other aspects of the problem of finding guaranteed error bounds for electronic structure problems. I considered clusters of eigenvalues, quantities of interest, localized basis as a discretization, and also studied 3D problems with numerical simulations on real 3D material systems.

2.2.2 *A posteriori* error bounds for linear eigenvalue problems

In this section, we consider the case $V_{\Phi_N}^{\mathrm{nl}} = 0$. We first present error estimations of an isolated eigenvalue, before moving to the estimation of a multiple or a cluster of eigenvalues.

Single eigenvalues in linear eigenvalue problems

The *a posteriori* error estimation for eigenvalue problems can be traced back at least to Kato, Temple [90, 141], Bauer and Fike [17]. They provided an *a posteriori* error bound on a matrix eigenvalue (not necessarily the lowest one) that depends on the residual of the approximate solution. For simplicity, and since the major part of this document deals with self-adjoint operators, we present these bounds in the case of an Hermitian matrix.

Theorem 2.2.1 (Bauer–Fike [17])**.** *Let ϕ_N, λ_N be an approximate eigenpair of an Hermitian matrix A with $\|\phi_N\|_2 = 1$, with residual*

$$\mathrm{Res} = A\phi_N - \lambda_N \phi_N.$$

There exists an eigenvalue λ of A such that

$$|\lambda_N - \lambda| \leq \|\mathrm{Res}\|_2.$$

This bound is very convenient for many reasons. First the bound depends on the 2-norm of the residual, which is easy to compute. Second, it does not involve constants. Moreover the proof is very short and can easily be extended to self-adjoint operators. As a drawback it gives a bound between the computed eigenvalue and *an* eigenvalue λ, which does not have to be the target eigenvalue. E.g. if λ_N aims at approximating the lowest eigenvalue, the corresponding λ may not be the lowest eigenvalue of the matrix A. More importantly it largely overestimates the error in practice, and does not converge to the error at the right rate, which means that the bound degrades when the approximation λ_N is improved. Also, it does not provide any error bound on the eigenvector error.

Kato and Temple have provided error bounds on the eigenvalues that improves the convergence speed by replacing the 2-norm by a 2-norm squared as follows.

Theorem 2.2.2 (Kato–Temple [90, 141]). *Let ϕ_N, λ_N be an approximate eigenpair of an Hermitian matrix A with $\|\phi_N\|_2 = 1$, such that $\lambda_N = \langle \phi_N, A\phi_N \rangle$, with residual*

$$\mathrm{Res} = A\phi_N - \lambda_N \phi_N.$$

Let λ be the eigenvalue closest to λ_N and δ the distance from λ_N to the rest of the spectrum, that is

$$\delta = \min_i \{|\lambda_i - \lambda_N|, \quad \lambda_i \neq \lambda, \quad \lambda \in \mathrm{Sp}(A)\}.$$

Then

$$|\lambda_N - \lambda| \leq \frac{\|\mathrm{Res}\|_2^2}{\delta}.$$

Compared to the Bauer–Fike bound, this bound is better because the norm of the residual is squared, hence when $\|\mathrm{Res}\|_2 \to 0$, the Kato–Temple bound is smaller than the Bauer–Fike bound. However, this bound exhibits a constant $1/\delta$ which is the gap between the eigenvalue of interest and the surrounding ones. This therefore requires that the approximate eigenvalue is simple. Moreover, when the gap is very small, the constant explodes. In practice, the bound is not optimal, in the sense that when the error decreases, the bound $\frac{\|\mathrm{Res}\|_2^2}{\delta}$ does not converge to the exact error $|\lambda_N - \lambda|$.

In practice, for electronic structure problems where the considered operator is $-\Delta + V$, what is observed is that the discretization error behaves asymptotically like

$$|\lambda_N - \lambda| \simeq \|A^{-1/2}\mathrm{Res}\|_2^2 = \langle \mathrm{Res}, A^{-1}\mathrm{Res} \rangle,$$

where Res is the residual for the approximate eigenpair (ϕ_N, λ_N). First, this means that the gap does not seem to play a role, at least when the discretization parameter goes to $+\infty$, to the error in the eigenvalues, and second, that the right norm to use is not the 2-norm but the dual norm of the residual, which requires inverting a linear system to evaluate. This was shown during my PhD thesis in [A8] for the Laplace eigenvalue problem. Therefore, efficiently estimating the dual norm of the residual is crucial to obtain bounds that are computable within a reasonable time.

From very simple to more involved but asymptotically exact, many bounds have been originally proposed. The earlier contributions include [61, 148, 18, 62, 108, 95, 96, 138, 67, 19, 120]. Some of the proposed bounds so far are specific to the estimation of the first eigenvalue, such as [9, 84, 85, 39, 150, 97, 135, 102]. For a review on the subject, see e.g. [112, 129] and references therein.

So far, we have only mentioned the error estimation for the eigenvalues. But in many applications, the error in the eigenvector is also important to control. Contributions to this are provided e.g. in [143, 103, 98, 83, 53, 71, 125]. However, many of these contributions contain uncomputable terms which are claimed of higher order on fine enough meshes in the finite element context via *a priori* arguments.

Finally, depending on the derived *a posteriori* error bounds, the discretization basis may be important. It especially plays a role in the estimation or evaluation of the dual norm of the residual. A typical approach in the articles [A10, A16] is therefore the following:

1. derive generic error bounds for the eigenvalues and eigenvectors that depend on (i)-the gap constant, and (ii)-the dual norm of the residual, such as

$$|\lambda - \lambda_N| \leq C\|A^{-1/2}\mathrm{Res}\|^2, \tag{2.2.1}$$

   where $C > 0$ depends on the gap,

2. estimate the gap depending on the context in order to estimate C,

3. estimate or evaluate the dual norm of the residual,

4. combine the steps 1., 2., 3. to obtain a fully computable and guaranteed upper bound of the error.

Step 3. is specific to the discretization basis. In general, for a Schrödinger operator, the dual norm of the residual can be bounded by the H^{-1}-norm of the residual upon some assumptions on the potential. E.g., if $V \geq 1$,

$$\|(-\Delta + V)^{-1/2}\mathrm{Res}\| \leq \|(-\Delta + 1)^{-1/2}\mathrm{Res}\|.$$

For a plane wave discretization, since the Laplace operator is diagonal in plane waves, as shown in [A10] the right-hand side is easily is computable. For the Laplace operator discretized with finite elements, techniques based on the resolution of local problems have been developed in [58] for a source problem and generalized to eigenvalue problems in [A10].

For a Schrödinger operator discretized with a localized basis, a recent work [A16] provides fully guaranteed and computable error bounds for localized basis sets (LCAOs), which we present in detail now. In this context, given parameters $\mathbf{R}_1, \ldots, \mathbf{R}_K \in \mathbb{R}^d$ corresponding to atomic positions, for $k = 1, \ldots, K$, let $V_k(x) = W(|x - \mathbf{R}_k|)$ with $W : \mathbb{R}^+ \to \mathbb{R}$ (typically $W(r) = 1/r$) and consider the linear Hamiltonian operator of Schrödinger-type defined by

$$A = -\frac{1}{2}\Delta + \sum_{k=1}^{K} V_k,$$

possibly shifted to ensure that the operator A is coercive. The exact and approximate (single) eigenvalues (not necessarily the lowest one) are respectively denoted by λ, λ_N and the exact and approximate eigenfunctions by ϕ, ϕ_N. The first step is to obtain an estimation of the error in the eigenvalues and eigenvectors that depends on the residual in the form of (2.2.1). The second step is to separate the residual into atomic components, using a partition of unity as illustrated in Figure 2.1 subordinate to a finite cover of $\mathbb{R}^d$ denoted by $(\Omega_k)_{1 \leq k \leq K+1}$, satisfying

$$p_k \in \mathrm{C}^\infty(\mathbb{R}^d),\ \mathrm{supp}(p_k) \subset \Omega_k,\ 0 \leq p_k \leq 1,\ \text{and } \forall x \in \mathbb{R}^d,\ \sum_{k=1}^{K+1} p_k(x) = 1, \tag{2.2.2}$$

where supp denotes the support. We also define atomic operators $A_k = -\frac{1}{2}\Delta + V_k$, for which computing the eigenvalues is simpler, and we estimate the dual norm of the residual by a sum of localized residuals over the subdomains Ω_k as

$$\|A^{-1/2}\mathrm{Res}\|^2 \leq \sum_{k=1}^{K+1} \|\sqrt{p_k}\,\mathrm{Res}\,\|^2_{A_k^{-1}}.$$

We are now left with estimating the terms on the right-hand side of the above equation. For this, we use a spectral decomposition of A_k which is easy to obtain as A_k is radially symmetric on a compact domain Ω_k. Let $(\varepsilon_j^{(k)}, \psi_j^{(k)}) \in \mathbb{R} \times \mathrm{H}^1_0(\Omega_k)$, $j \in \mathbb{N}$, be the eigenpairs of A_k such that $0 < \varepsilon_1^{(k)} \leq \varepsilon_2^{(k)} \leq \ldots$ counting multiplicities and

$$A_k\psi_j^{(k)} = \varepsilon_j^{(k)}\psi_j^{(k)}, \quad \text{with} \quad \langle \psi_j^{(k)}, \psi_{j'}^{(k)} \rangle_{\Omega_k} = \delta_{jj'}, \quad j, j' \geq 1.$$

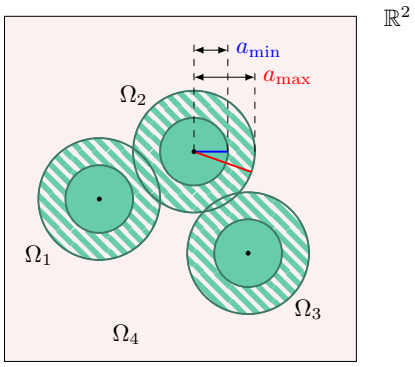


Figure 2.1: The case of three nuclei in $\mathbb{R}^2$. Finite cover composed of three balls $\Omega_{1,2,3}$, of radius $r = a_{\max}$ centered at nuclei, and the unbounded set Ω_4 (in light red). The hashed region is the overlap between sets of the cover. For $k = 1, 2, 3$, the partition functions p_k take the value one on the smaller balls of radius $r = a_{\min}$ (rigid green). On $\mathbb{R}^2 \setminus (\Omega_1 \cup \Omega_2 \cup \Omega_3)$, p_4 is equal to 1. Otherwise, in the hashed region, p_k is strictly between 0 and 1.

By spectral calculus, we can write

$$\|v\|^2_{A_k^{-1}} = \sum_{j=1}^{\infty} \frac{1}{\varepsilon_j^{(k)}} \left| \langle v, \psi_j^{(k)} \rangle_{\Omega_k} \right|^2 .$$

As this sum is *a priori* not computable, we introduce the partial dual norm associated to the partial sum of finite degree $J_k \in \mathbb{N}$ defined by

$$\mathcal{I}_k(v) := \sum_{j=1}^{J_k} \frac{1}{\varepsilon_j^{(k)}} \left| \langle v, \psi_j^{(k)} \rangle_{\Omega_k} \right|^2 .$$

We estimate the dual norm of the residual by the following computable quantity as presented in [A16, Theorem 3.9]:

$$\|A^{-1/2}\mathrm{Res}\|^2 \leq \sum_{k=1}^{K+1} \left[\mathcal{I}_k(\sqrt{p_k}\,\mathrm{Res}) + \frac{1}{\varepsilon_{J_{k+1}}^{(k)}} \left(\|\sqrt{p_k}\,\mathrm{Res}\,\|^2_{\Omega_k} - \sum_{j=1}^{J_k} \left| \langle \sqrt{p_k}\,\mathrm{Res}, \psi_j^{(k)} \rangle_{\Omega_k} \right|^2 \right) \right] . \tag{2.2.3}$$

This estimation combined with (2.2.1) gives rise to very satisfactory error bounds that are shown in Figure 2.2 both on the eigenvalue and the eigenvectors. The parameter l corresponds to the chosen overlap in the partition of unity, and has a large impact on the quality of the error bounds.

We then used this error bound to adaptively enlarge the discretization basis. For any atom indexed by $1 \leq k \leq K$, we define the local error indicator by

$$\eta_k^2 := \mathcal{I}_k(\sqrt{p_k}\,\mathrm{Res}) + \frac{1}{\varepsilon_{J_k+1}^{(k)}} \left(\|\sqrt{p_k}\,\mathrm{Res}\,\|^2_{\Omega_k} - \sum_{j=1}^{J_k} \left| \langle \sqrt{p_k}\,\mathrm{Res}, \psi_j^{(k)} \rangle_{\Omega_k} \right|^2 \right) . \tag{2.2.4}$$

The refinement strategy then consists in finding for which atom k the local error indicator η_k is the largest and increasing the number of basis functions centered on the k^{th} atom.

We numerically observe that for a system (in 1D) with two different charges, the adaptive strategy yields a nice convergence rate improvement compared to putting the same number of

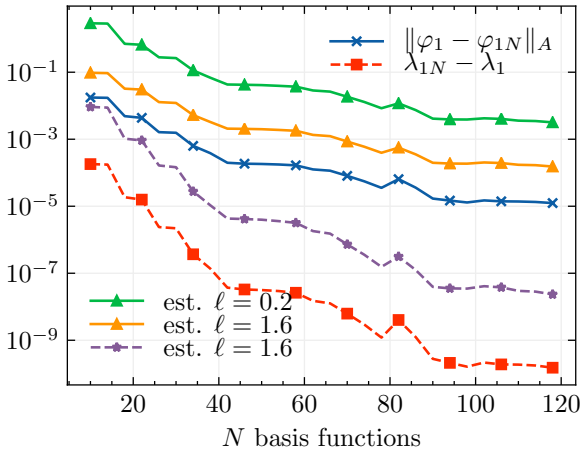


Figure 2.2: Estimates of errors $\|\phi_1 - \phi_{1N}\|_A$ (solid lines) and $\lambda_{1N} - \lambda_1$ (dashed line) for first eigenpair (λ_1, ϕ_1) of (1.2.7) with $V^{\mathrm{nl}}_{\Phi_N} = 0$, using (2.2.3). The physical system contains two atoms in 1D. On the x-axis is the number of AO basis functions equal to $N = n_1 + n_2$ with $n_1 = n_2 = N/2$ per atom.

degrees of freedom on each atom, as shown on the left of Figure 2.3. Moreover for a system with two identical charges, which is shown on the right of Figure 2.3, we observe that the number of basis functions stays the same on each atom, as expected.

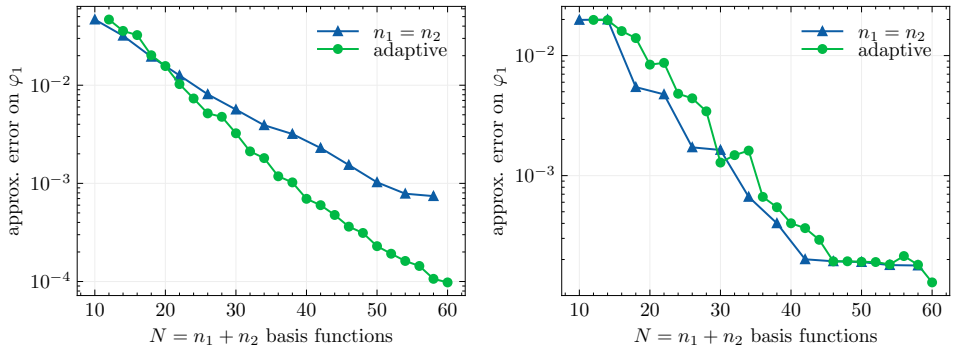


Figure 2.3: Left: distinct nuclear charges $z_{-R} = 1$, $z_R = 3.$; Right: identical nuclear charges $z_{-R} = z_R = 1$.Adaptive basis sets on molecules in 1D. Exact operator norm error $\|\phi_1 - \phi_{1N}\|_A$ for adaptive basis set (in green) versus uniformly refined one with $n_1 = n_2$ (in blue).

To summarize the error estimation exploits the sharp decrease of the atomic potential, and the fact that atomic problems can be more easily solved. The partition of unity enables the localization of the error that is exploited to adaptively refine the discretization basis.

Multiple and clustered eigenvalues

In many contexts, several eigenvalues have to be computed together with their corresponding eigenvectors. As mentioned in the previous section, this causes problems as the gap between eigenvalues is potentially zero, making the single-eigenvalue error bounds not relevant as they explode. In this case we consider the density matrix, i.e. the orthogonal projector on the eigenvectors, which is well-defined when there is a gap between the considered eigenvalues and the surrounding eigenvalues.

Several works so far have been dealing with the approximation of clusters of eigenvalues. Depending on the contributions, the authors sometimes consider the error as the distance from the computed eigenvector to the space spanned by the set of exact eigenfunctions [71, 66]. In this case, the norms to consider are the usual Sobolev norms, depending on the considered differential operator. E.g. in the case of the Laplace operator, the natural norm for the eigenvectors is the Sobolev H^1 norm. In some other cases, in particular the approach developed in [A10], the density matrix error is directly estimated. In that case, the norms to consider are Hilbert–Schmidt norms denoted by $\mathfrak{S}_2(\mathcal{H})$ below adapted to the operators at stake in the equations. The functional analysis setting is precisely described in [A10, Section 2.2].

The obtained error bounds are valid under the crucial assumption that there is a gap between the cluster of considered eigenvalues (from index p to P counting multiplicities) and surrounding eigenvalues, as follows. We denote with a subscript N approximate quantities.

Assumption 2.2.3 (Gap assumption)**.** *There holds $\lambda_{p-1} < \lambda_p$ if $p > 1$ and $\lambda_P < \lambda_{P+1}$, and there exist $\underline{\lambda}_{P+1}$, and $\overline{\lambda}_{p-1}$ if $p > 1$, such that there holds*

$$\lambda_{p-1} \leq \overline{\lambda}_{p-1} < \lambda_{pN} \text{ when } p > 1, \qquad \lambda_{PN} < \underline{\lambda}_{P+1} \leq \lambda_{P+1}.$$

The first result to highlight here is a theorem linking the error in the density matrices denoted by γ and the sum of the eigenvalues in the considered cluster.

Theorem 2.2.4 (Eigenvalue bounds)**.** *Let Assumption 2.2.3 hold and let the density matrices γ^0 and γ_N be respectively the exact and approximate density matrices. Then*

$$\|A^{1/2}(\gamma^0 - \gamma_N)\|^2_{\mathfrak{S}_2(\mathcal{H})} - \lambda_P \|\gamma^0 - \gamma_N\|^2_{\mathfrak{S}_2(\mathcal{H})} \leq \sum_{i=p}^{P} (\lambda_{iN} - \lambda_i) \leq \|A^{1/2}(\gamma^0 - \gamma_N)\|^2_{\mathfrak{S}_2(\mathcal{H})}. \quad (2.2.5)$$

This means that bounds on the eigenvalues and the eigenvectors can be obtained similarly. For conciseness, we only present the obtained error bounds on the eigenvalues. Another interesting property is that the errors on density matrices $\|A^{1/2}(\gamma^0 - \gamma_N)\|^2_{\mathfrak{S}_2(\mathcal{H})}$ and $\|\gamma^0 - \gamma_N\|^2_{\mathfrak{S}_2(\mathcal{H})}$ can be efficiently computed from the eigenfunctions. Provided that these conditions can be verified, we obtain a guaranteed error bound on the eigenvalues (and hence the density matrices) as stated in the following theorem [A10, Theorem 5.9].

Theorem 2.2.5 (Guaranteed bounds for the sum of eigenvalues)**.** *Let $p, P \in \mathbb{N}\backslash\{0\}, P \geq p$. For $i = p, \ldots, P$, let $(\phi_{iN}, \lambda_{iN}) \in \mathcal{V}_N \times \mathbb{R}_+$ be an eigenpair solution to equation* (1.2.7) *discretized with plane waves. Let $\underline{\lambda}_{P+1}$, and $\overline{\lambda}_{p-1}$ if $p > 1$ satisfying Assumption 2.2.3. For $i = p, \ldots, P$, define $\mathrm{Res}_N = A\phi_{iN} - \lambda_{iN}\phi_{iN}$. Then*

$$0 \leq \sum_{i=p}^{P} (\lambda_{iN} - \lambda_i) \leq (1 + c_N) \sum_{i=p}^{P} \|\mathrm{Res}_N\|^2_{H^{-1}(\Omega)}, \quad (2.2.6)$$

with

$$c_N = \frac{c\lambda_{PN}}{N^2} \max\left[\left(\frac{\lambda_{pN}}{\overline{\lambda}_{p-1}} - 1\right)^{-1}, \left(1 - \frac{\lambda_{PN}}{\underline{\lambda}_{P+1}}\right)^{-1}\right]^2, \quad (2.2.7)$$

with a computable $c > 0$.

The bound on the eigenvalues mainly depends on the sum of the individual residuals, which can be computed independently, and in parallel if needed. The remaining term c_N in the prefactor is (i) computable, (ii) decreases to 0 when N goes to infinity. Numerical results illustrating the error bound in the case of a 1D Schrödinger operator with periodic boundary conditions is presented in Figure 2.4. We indeed observe that the error on the eigenvalues (in red) is closely approximated by the proposed error bound (in blue) and when the number of degrees of freedom (ndof) is large, the actual error and the error bounds match. A similar result has been obtained for a finite element discretization.

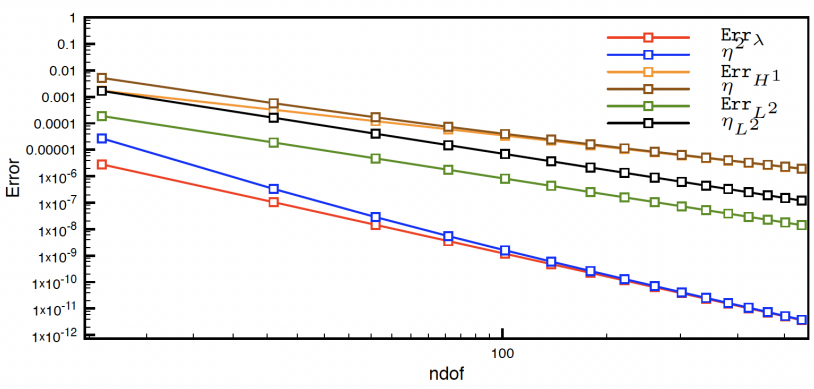


Figure 2.4: Convergence of the errors and their upper bounds for a 1D Schrödinger operator with periodic boundary conditions with $p = 2$, $P = 3$.

2.2.3 *A posteriori* error estimation for nonlinear eigenvalue problems

In the case of nonlinear eigenvalue problems as (1.1.4), an extra difficulty comes from the nonlinear term $V_\Phi^{\mathrm{nl}} \neq 0$. So far, available fully guaranteed and computable error bounds hold in the case where the energy functional in (1.2.1) is convex. To my knowledge, no global guaranteed bound has been derived in nonconvex cases. However, in practice, one often observes that the error bounds remain valid outside of the regime where they are guaranteed, and can provide practically relevant error bounds.

A review of different error bounds available in this case is presented in [A20]. The main references in the field are the following. In [104], an error bound is provided for the Hartree–Fock model, that is not guaranteed, but seems to work well in practice, even for local basis set discretizations. In [A19], a fully guaranteed error bound is provided for the Gross–Pitaevskii equation, which consists in finding the single lowest eigenvalue and the corresponding eigenvector of a nonlinear operator where the nonlinear potential takes the simple form $V_\Phi^{\mathrm{nl}} = |\Phi|^2$ and is convex. In this case, the sought-for eigenvalue is well separated from the rest of the spectrum. The error bound is fully guaranteed and relies on the use of Theorem 2.1.1 for the estimation. It was recently extended in [A3] to the case of several eigenvalues with a convex energy functional. In [A5] practical error bounds for nonconvex problems are derived. They are not fully guaranteed but turn out to work well in practice, as will be shown in the next section. Another work [132] is not directly related to eigenvalue problems but uses a similar theory as in Theorem 2.1.1 and could in principle be also used for nonlinear eigenvalue problems. Other works focus on finding error bounds or building adaptive meshes such as [48, 40, 47] but the bounds are in general not computable, that is the unknown constants are not considered in practice.

We now provide some detail about the results obtained in [A3] in the case of a nonlinear operator for which the energy functional is convex: the energy is a functional of the density written as

$$E(\gamma) \coloneqq \mathrm{Tr}(A\gamma) + F(\rho_\gamma), \tag{2.2.8}$$

where F is convex on $\mathcal{M}$ defined in (1.2.2) and taken smooth enough so that for each $\gamma \in \mathcal{M}$ there exists $V_{\rho_\gamma} \in L^2_{\mathrm{per}}(\Omega;\mathbb{R})$ such that, for any $\widetilde{\gamma} \in \mathcal{M}$,

$$\langle F'(\rho_\gamma), \rho_{\widetilde{\gamma}} \rangle = \int_\Omega V_{\rho_\gamma}\rho_{\widetilde{\gamma}} = \mathrm{tr}\big(V_{\rho_\gamma}\widetilde{\gamma}\big), \tag{2.2.9}$$

with a slight abuse of notation for the last term of the previous equation. The eigenvalue

problem (1.2.7) can be recast in this case as: Find $\Phi_N = (\phi_{1,N}, \ldots, \phi_{N_{\mathrm{el}},N}) \in \mathcal{V}_N$, $\Lambda_N = (\lambda_{1,N}, \ldots, \lambda_{N_{\mathrm{el}},N}) \in \mathbb{R}^{N_{\mathrm{el}}}$ satisfying for $i, j = 1, \ldots, N_{\mathrm{el}}$,

$$\begin{aligned} \left(A + V^{\mathrm{nl}}_{\rho_{\gamma_N}}\right) \phi_{i,N} &= \lambda_{i,N} \phi_{i,N}, \\ \langle \phi_{i,N}, \phi_{j,N} \rangle &= \delta_{ij}, \\ \gamma_N &= \sum_{i=1}^{N_{\mathrm{el}}} |\phi_{j,N}\rangle\langle\phi_{i,N}|. \end{aligned} \tag{2.2.10}$$

The following simple calculation is key in the analysis to obtain guaranteed error bounds in the nonlinear case: For any $\mu \in \mathbb{R}$,

$$\begin{aligned} E(\gamma_N) - E(\gamma^0) = \operatorname{tr}\left((A + V_{\rho_{\gamma_N}} - \mu)\gamma_N\right) - \operatorname{tr}\left((A + V_{\rho_{\gamma_N}} - \mu)\gamma^0\right) \\ - \left(F(\rho_{\gamma^0}) - F(\rho_{\gamma_N}) - \langle F'(\rho_{\gamma_N}), \rho_{\gamma^0} - \rho_{\gamma_N} \rangle\right) \ . \end{aligned} \tag{2.2.11}$$

Now since the functional F is convex, the term $-\left(F(\rho_{\gamma^0}) - F(\rho_{\gamma_N}) - \langle F'(\rho_{\gamma_N}), \rho_{\gamma^0} - \rho_{\gamma_N} \rangle\right)$ is negative and can therefore be dropped in the estimation of the energy. Second, the estimation being valid for any $\mu \in \mathbb{R}$, one can choose the best $\mu \in \mathbb{R}$ to make the right-hand side of the previous equation as small as possible, but still positive and computable. A good $\mu \in \mathbb{R}$ is such that $-\operatorname{tr}\left((A + V_{\rho_{\gamma_N}} - \mu)\gamma^0\right)$ is negative, so that the upper bound

$$E(\gamma_N) - E(\gamma^0) \leq \operatorname{tr}\left((A + V_{\rho_{\gamma_N}} - \mu)\gamma_N\right),$$

does not depend on γ^0 anymore. Moreover, the larger μ is, the smaller is the bound. Therefore, one has to find a μ as large as possible satisfying

$$\mu \leq \frac{1}{N_{\mathrm{el}}} \sum_{i=1}^{N_{\mathrm{el}}} \varepsilon_i,$$

where $\varepsilon_1, \ldots, \varepsilon_{N_{\mathrm{el}}}$ are the lowest eigenvalues of the operator $A + V_{\rho_{\gamma_N}}$. This requires a lower bound of the mean of the eigenvalues of a linear eigenvalue problem.

As described in Chapter 1, for a fixed discretization parameter N, the density matrix γ_N solution to (2.2.10) is not directly computable but is obtained as the limit of a sequence $(\gamma_{N,m})_{m \in \mathbb{N}}$ generated by SCF iterations (see Section 1.2.3). Hence, we replace γ_N by $\gamma_{N,m}$ from (1.2.14), and $\varepsilon_i = \lambda_{i,N,m}$. In other words, we compute $\mu^{\mathrm{lb}}_{N,m+1}$ as a lower bound of the mean of the eigenvalues from the (infinite-dimensional) eigenvalue problem (2.2.10) with $A = A + V^{\mathrm{nl}}_{\rho_{\gamma_{N,m}}}$. The bound for the mean of the lowest eigenvalues can e.g. be based on the guaranteed error estimates for eigenvalue clusters [A10]. In [A3] we also propose more accurate, although not fully guaranteed, error estimates which are easily computable. Having the needed lower bound $\mu^{\mathrm{lb}}_{N,m+1}$, one obtains a bound for the energy error as

$$0 \leq E(\gamma_{N,m}) - E(\gamma^0) \leq \operatorname{tr}\left((A + V_{\rho_{\gamma_{N,m}}} - \mu^{\mathrm{lb}}_{N,m+1})\gamma_{N,m}\right), \tag{2.2.12}$$

and yields the following certification of the energy at iteration m of the SCF algorithm:

$$E(\gamma^0) \in \left[E(\gamma_{N,m}) - \operatorname{tr}\left((A + V_{\rho_{\gamma_{N,m}}} - \mu^{\mathrm{lb}}_{N,m+1})\gamma_{N,m}\right), E(\gamma_{N,m})\right]. \tag{2.2.13}$$

Informally, this means that we can control the error at iteration m via doing an extra iteration of the SCF algorithm.

Separating between discretization and SCF error We can then use these bounds to adapt the parameters of the simulation and obtain optimized parameters (discretization and number of iterations) to reach a given accuracy. Indeed we separate the error bound (2.2.12) into two parts: one depending mainly on the discretization

$$\mathtt{err}_{N,m}^{\mathrm{disc}} := \mathrm{tr}\left((A + V_{\rho_{\gamma_{N,m}}} - \mu_{N,m+1}^{\mathrm{lb}})\gamma_{N,m+1}\right) \geq 0, \tag{2.2.14}$$

and the other depending mainly on the number of SCF iterations

$$\mathtt{err}_{N,m}^{\mathrm{SCF}} := \mathrm{tr}\Big((A + V_{\rho_{\gamma_{N,m}}})\gamma_{N,m}\Big) - \mathrm{tr}\Big((A + V_{\rho_{\gamma_{N,m}}})\gamma_{N,m+1}\Big) \geq 0. \tag{2.2.15}$$

Then, as $\mathrm{tr}(\gamma_{N,m+1}) = \mathrm{tr}(\gamma_{N,m})$, we naturally have that

$$\mathrm{tr}\left((A + V_{\rho_{\gamma_{N,m}}} - \mu_{N,m+1}^{\mathrm{lb}})\gamma_{N,m}\right) = \mathtt{err}_{N,m}^{\mathrm{disc}} + \mathtt{err}_{N,m}^{\mathrm{SCF}}.$$

In [A3] we proved the following theorem, providing a fully guaranteed and computable bound of the energy error, which, moreover, separates relative to the different sources of errors.

Theorem 2.2.6 (Fully guaranteed error bound on the energy)**.** *Under some assumptions, there holds*

$$E(\gamma_{N,m}) - E(\gamma^0) \leq \mathtt{err}_{N,m}^{\mathrm{disc}} + \mathtt{err}_{N,m}^{\mathrm{SCF}}, \tag{2.2.16}$$

where $\mathtt{err}_{N,m}^{\mathrm{disc}}$ *and* $\mathtt{err}_{N,m}^{\mathrm{SCF}}$ *are respectively defined by* (2.2.14) *and* (2.2.15)*.*

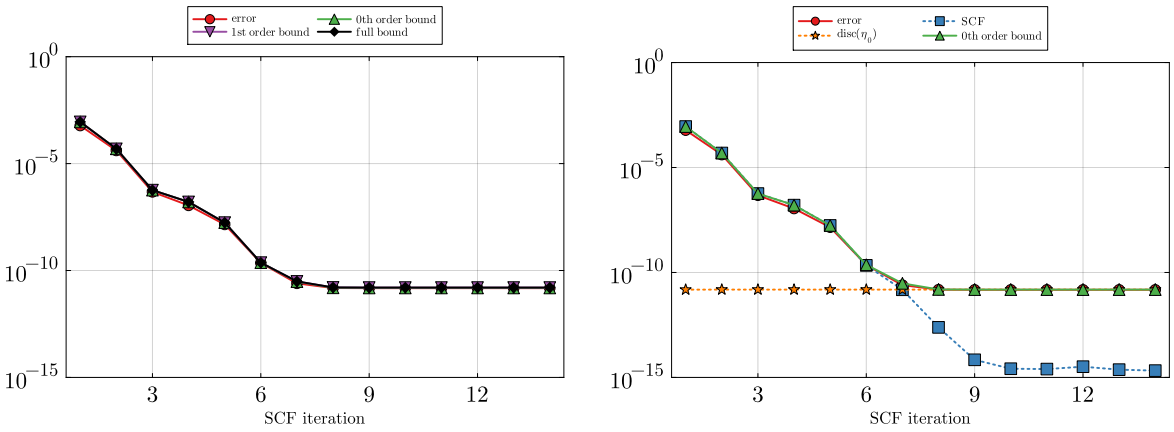


Figure 2.5: Tracking of the error $E(\gamma_{N,m}) - E(\gamma^0)$ for a Si crystal (eight $\boldsymbol{k}$-points). (Left) Full-inversion bound with zeroth- and first-order approximations. (Right) Zeroth-order bound and its splitting between SCF and discretization contributions, as in (2.2.16).

We remark that (i) $\mathtt{err}_{N,m}^{\mathrm{SCF}}$ goes to zero as the number of SCF iterations m goes to infinity, provided that the SCF algorithm does converge, and (ii) $\mathtt{err}_{N,m}^{\mathrm{disc}}$ tends to zero as the discretization space is enlarged, provided that the limit of $\mathcal{V}_N$ when $N \to +\infty$ is $H^1_{\mathrm{per}}(\Omega)$ in the sense that $\forall \phi \in H^1_{\mathrm{per}}(\Omega)$, $\|\phi - \Pi_N \phi_{H^1_{\mathrm{per}}(\Omega)}\| \to 0$ as $N \to \infty$, and that $\mu_{N,m+1}^{\mathrm{lb}}$ is well chosen, for instance as explained above. This method is illustrated on a series of test cases. For the purpose of this manuscript, we present in Figure 2.5 one of these results in a Silicon crystal. On the left part of the figure, we plot the energy error, and the corresponding error bounds computed with several methods (either with 0^{th} order or 1^{st} order approximation of the dual norm of the residual – cheap to compute but not fully guaranteed, or with the full inversion –

more expensive but fully guaranteed). We observe that all the bounds closely match the error in the energy. On the right part of the figure, we compare the different components of the error bound when separated between discretization and SCF error. We observe that, as predicted, the SCF error decreases down to machine precision, while the discretization error is constant (since the discretization is fixed in this example) and becomes the largest part of the error at the end of the iterations.

2.2.4 *A posteriori* error estimation for quantities of interest

Once error bounds are available on the eigenvalues and the eigenvectors and/or the density matrix, it is natural to wonder how these bounds can be extended for quantities of interest. This has been investigated in [A5] for interatomic forces. Let us first explain the main ideas and the difficulties that are faced when studying quantities of interest.

Assume we want to find $x \in \mathbb{R}^n$ such that $F(x) = 0$, for some nonlinear function $F : \mathbb{R}^n \to \mathbb{R}^n$ (the residual). In this contribution, the first task was to find the Jacobian DF_x, which is not trivial due to degeneracies (eigenvalues e.g.) and constraints (the density matrix has to belong to $\mathcal{M}$ defined in (1.2.2)) of the input space. For this, we strongly used the geometrical framework developed in [38]. Second, the computation of the Jacobian is very expensive in this context, so that an iterative procedure has been developed to limit the computational cost of the Jacobian. Third, choosing the norm to compute the errors and the residual is not obvious, as mentioned in the beginning of this chapter. Otherwise, the errors can be largely overestimated. As in other works, we considered Sobolev norms, with the aim of making DF a bounded operator between the relevant function spaces.

The main result of this work therefore lies in the derivation of an efficient, asymptotically accurate, way of approximating $\nabla Q(x) \cdot \left(DF_x^{-1} F(x)\right)$ using the specific structure of the residual $F(x)$ in a plane wave discretization, where Q represents a component of the interatomic forces of the system. This approximation can then be used either to approach the actual error $Q(x) - Q(x_*)$ in (2.1.5) or to improve $Q(x)$ by computing $Q(x) - \nabla Q(x) \cdot \left(DF_x^{-1} F(x)\right)$, which is a better approximation of $Q(x_*)$.

To do so, we leverage the fact that the Laplace operator is diagonal in plane waves. We decompose the error on the low-frequency and high-frequency modes, and use a Schur complement to improve the error bound, while limiting its computational cost. We denote by $F_{\mathrm{Schur}}(\gamma)$ the new residual satisfying

$$F_{\mathrm{Schur}}(\gamma) = \begin{bmatrix} (\boldsymbol{\Omega} + \boldsymbol{K})_{11}^{-1} \left(F_1 - (\boldsymbol{\Omega} + \boldsymbol{K})_{12} \boldsymbol{M}_{22}^{-1} F_2\right) \\ \boldsymbol{M}_{22}^{-1} F_2 \end{bmatrix},$$

where $\boldsymbol{M}_{22}$ is substantially the Laplace operator, so that it is very easy to invert. The operator $(\boldsymbol{\Omega} + \boldsymbol{K})_{11}$ only involves low-frequencies.

In Figure 2.6, we plot the new estimate $\mathrm{d}Q(\gamma) \cdot (R_{\mathrm{Schur}}(\gamma))$ of the error $Q(\gamma) - Q_*$ as well as the differences

$$\begin{aligned} Q_{\mathrm{err}} - Q_* &\coloneqq Q(\gamma) - \mathrm{d}Q(\gamma) \cdot (\boldsymbol{\Pi}_\gamma(\gamma - \gamma_*)) - Q_*, \\ Q_{\mathrm{res}} - Q_* &\coloneqq Q(\gamma) - \mathrm{d}Q(\gamma) \cdot (\boldsymbol{M}^{-1} R(\gamma)) - Q_*, \\ Q_{\mathrm{Schur}} - Q_* &\coloneqq Q(\gamma) - \mathrm{d}Q(\gamma) \cdot (R_{\mathrm{Schur}}(\gamma)) - Q_*, \end{aligned}$$

in order to have a better estimation of the improvement on the estimation of the error. With the Schur complement method, the new estimate better matches the error than the crude one simply using the residual: the accuracy of the estimation is approximately improved by one order of magnitude.

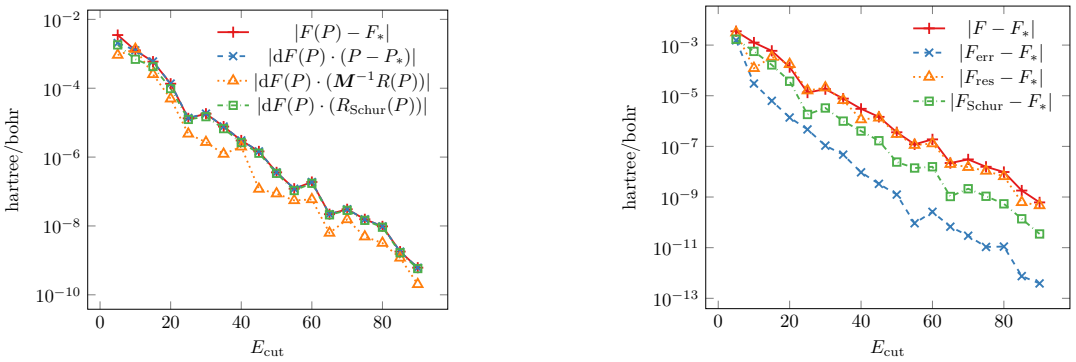


Figure 2.6: (Left) Estimation of the error $Q(\gamma) - Q_*$ with $\mathrm{d}Q(\gamma) \cdot X$ where X is either the exact error $\mathbf{\Pi}_\gamma(\gamma - \gamma_*)$, the preconditioned residual $\boldsymbol{M}^{-1}R(\gamma)$ or the modified residual $F_{\text{Schur}}(\gamma)$. (Right) Enhancement of the estimation of the forces by replacing $Q(\gamma)$ with $Q(\gamma) - \mathrm{d}Q(\gamma) \cdot X$ where X is either the exact error $\mathbf{\Pi}_\gamma(\gamma - \gamma_*)$, the preconditioned residual $\boldsymbol{M}^{-1}F(\gamma)$ or the modified residual $F_{\text{Schur}}(\gamma)$.

To conclude, obtaining error bounds on quantities of interest is possible, but requires a special care about the chosen norms and estimations, and has to take into account the structure of the problem.

2.3 Perturbation theory and beyond

In this section, we turn to *a priori* error estimates based on perturbation theory. The essence of perturbation theory is to provide Taylor expansions of quantities of interest such as eigenvalues or eigenfunctions, given the Taylor expansion of a differential operator. The obtained expressions can often be used to compute improved solutions at a low computational cost similarly to (2.1.3), or provide with information on convergence rates with respect to some parameter of the model. We present this theory with different flavors in this Section. We first present standard perturbation theory in the context of eigenvalue problems, which is extensively presented in Kato's book [91] in Section 2.3.1. We then present the contribution [A17] which extends perturbation theory in a context where one does not have a single reference problem but several. Third, we present in Section 2.3.3 another object called Feshbach–Schur map where the transformation between a reference problem and the problem of interest writes as a Schur complement.

2.3.1 Perturbation theory

The rationale behind perturbation theory is to find an approximation of an eigenvalue problem of interest by relating it to a known reference problem. In this section, we consider a reference Hamiltonian denoted by H^0 that is self-adjoint in a given Hilbert space, and we consider a family of Hamiltonian operators depending on some operator G

$$H_G := H^0 + G.$$

We consider admissible G's as a set of self-adjoint operators that are H^0-form bounded. This mainly imposes that the essential spectrum of H_G is the same as that of H^0. For simplification of the presentation, we only deal with the case of a single (lowest) eigenvalue λ_G, where, moreover, the separation between the considered eigenvalue and the rest of the spectrum is lower bounded over the set of admissible G's. However, this can easily be extended to multiple and

clusters of eigenvalues (see [126]). The orthogonal projector on the corresponding eigenvalue is denoted by γ_G and can be defined for any G using the resolvent $(z - H_G)^{-1}$ as the contour integral

$$\gamma_G := \frac{1}{2\pi i} \oint_{\mathcal{C}} (z - H_G)^{-1} \, \mathrm{d}z,$$

where the contour $\mathcal{C}$ can be the same for all G's due to the assumption on the admissible set of operators. For deriving the perturbation theory formulas, we also need to define the pseudo-inverse operator

$$K_G := \begin{cases} \left((\lambda_G - H_G)_{|(\mathrm{Ker}(\lambda_G - H_G))^\perp} \right)^{-1} & \text{on } (\mathrm{Ker}\,(\lambda_G - H_G))^\perp, \\ 0 & \text{on } \mathrm{Ker}\,(\lambda_G - H_G). \end{cases} \tag{2.3.1}$$

The reference Hamiltonian is taken as H_{G_1} for some G_1 so for this operator, we assume that we know the eigenvalue λ_{G_1} and a corresponding eigenvector exactly.

Standard perturbation theory consists in providing an approximation of γ_G as a series in the difference $H_G - H_{G_1}$ which in this context is just $G - G_1$. It relies on the widely known resolvent identity

$$(z - H_G)^{-1} = (z - H_{G_1})^{-1} \left(1 - (G - G_1)\,(z - H_{G_1})^{-1} \right)^{-1}, \tag{2.3.2}$$

and a Neumann expansion of the operator $\left(1 - (G - G_1)\,(z - H_{G_1})^{-1} \right)^{-1}$. Integrating over the chosen contour in order to obtain the orthogonal projector on the eigenvector, it is natural to define for $n \in \mathbb{N}$,

$$\mathcal{T}_n := \frac{1}{2\pi i} \oint_{\mathcal{C}} (z - H_{G_1})^{-1} \left((G - G_1)(z - H_{G_1})^{-1} \right)^n \mathrm{d}z, \qquad \gamma_n := \sum_{p=0}^{n} \mathcal{T}_p,$$

which are the different terms appearing in the Neumann expansion.

The following classical result [91, 126] relates the error between the exact orthogonal projector γ_G and its approximation at order n to the difference in the operators. A complete version can be found in [A16, Proposition 2.2].

Proposition 2.3.1 (Standard linear perturbation theory bound)**.** *Let G_1, G be two operators defined as above. There exists a constant $c > 0$ independent of $G - G_1$ such that for any $n \in \mathbb{N}$,*

$$\|\gamma_G - \gamma_n\| \leq c\,(C\,\|G - G_1\|)^{n+1}, \tag{2.3.3}$$

with $C := \max_{z \in \mathcal{C}} \left\| (z - H_{G_1})^{-1} \right\|$.

The operator norms appearing in (2.3.3) are standard operator norms. Note that other norms can be considered e.g. operator norm linked to H^0 for the projectors, in which case the corresponding dual norm has to be used for measuring the difference $G - G_1$. The terms in γ_n can be obtained explicitly using Cauchy's integration formula. For instance, $\mathcal{T}_0$ and $\mathcal{T}_1$ read

$$\begin{aligned} \mathcal{T}_0 &= \gamma_{G_1}, \\ \mathcal{T}_1 &= \gamma_{G_1}(G - G_1)K_{G_1} + K_{G_1}(G - G_1)\gamma_{G_1}. \end{aligned}$$

Therefore, perturbation theory provides an efficient way to approximate the orthogonal projector of an operator that is close to a well-known operator, which, in the context of electronic structure calculations, is typically the Laplace operator. Nevertheless, one can also imagine using perturbation theory in the case where the nuclei positions slightly move and use the orthogonal projector for a given configuration $\mathbf{R}$ as the base for a different configuration. This will actually be the spirit of the Grassmann extrapolation method presented in Section 4.2, however, with several reference points, which brings us to introduce multipoint perturbation theory.

2.3.2 Multipoint perturbation theory

As presented in [A17], a natural extension of single-point perturbation theory is to consider several reference operators and to extend the resolvent identity (2.3.2) to leverage several reference operators. We denote these operators by $G_1, G_2, \dots, G_m \in \mathcal{G}$, where $\mathcal{G}$ is the space of admissible operators G's. Therefore we are interested in the case

$$G \approx \sum_{i=1}^{m} \alpha_i G_i.$$

To simplify, we restrict ourselves to the case $\sum_{i=1}^{m} \alpha_i = 1$. The next theorem which is a simplified version of [A17, Theorem 3.1] provides such extended resolvent formula: we link the resolvent of a given G to the resolvent of the G_i's.

Theorem 2.3.1 (Multipoint resolvent formula)**.** *Let $\boldsymbol{\alpha} = (\alpha_j)_{j=1}^m \in \mathbb{R}^m$, $\boldsymbol{G} = (G_j)_{j=1}^m \in \mathcal{G}^n$, $G \in \mathcal{G}$, $z \in \mathbb{C} \backslash \left(\sigma(H_G) \cup_{j=1}^m \sigma(H_{G_j})\right)$. We define the operators*

$$\mathbb{H}_z := \sum_{1 \le i < j \le n} \alpha_i \alpha_j \left(G_i - G_j\right) \left(z - H_{G_j}\right)^{-1} \left(G_i - G_j\right) \left(z - H_{G_i}\right)^{-1},$$
$$\mathbb{L}_z := \sum_{j=1}^{m} \alpha_j \left(z - H_{G_j}\right)^{-1}.$$

We also define $W := G - \sum_{j=1}^m \alpha_j G_j$. If $1 + \mathbb{H}_z - W\mathbb{L}_z$ is invertible, there holds

$$(z - H_G)^{-1} = \mathbb{L}_z \left(1 + \mathbb{H}_z - W\mathbb{L}_z\right)^{-1}. \tag{2.3.4}$$

This theorem shows that the resolvent $(z - H_G)^{-1}$ can be expressed as a linear combination $\mathbb{L}_z$ of the resolvents of G_i's, up to an operator $(1 + \cdot)^{-1}$, as in the standard perturbation case. Therefore, compared to standard perturbation theory, the single resolvent is replaced by a linear combination of resolvents, and the remaining term $(G - G_1)(z - H_{G_1})^{-1}$ is replaced by a sum of two terms, contributing for two different errors. The term $-W\mathbb{L}_z$ is similar to the remaining term in standard perturbation theory $(G - G_1)(z - H_{G_1})^{-1}$. The additional term $\mathbb{H}_z$ is the most interesting one, as is contains the error terms $G_i - G_j$ twice, and is for this reason of higher order. This extension seems very natural, but to my knowledge, has not been previously presented elsewhere.

Interestingly, this resolvent formula can be transformed, as in the standard perturbation theory bound as an expansion giving an approximation of the resolvent order by order. This expansion can then be transformed into an expansion on the orthogonal projector γ_G by integrating along the considered complex contour using Cauchy's formula. Defining γ_n as the approximation at order n in $G_i - G_j$ of γ_G using the multipoint perturbation formula (2.3.4), we can obtain a bound comparable to what is obtained in standard perturbation theory.

Corollary 2.3.2 (Multipoint perturbation bound)**.** *Let $(G_j)_{j=1}^m \in \mathcal{G}^m$, and γ_n be the multipoint approximation at order n. There holds*

$$\|\gamma_G - \gamma_n\| \lesssim \delta_{\boldsymbol{\alpha},\boldsymbol{G}}^{n+1+\xi_n} + \delta_W^{n+1}, \tag{2.3.5}$$

with

$$\delta_{\boldsymbol{\alpha},\boldsymbol{G}} := \max_{1 \le i,j \le m} \sqrt{|\alpha_i \alpha_j|}\, \|G_i - G_j\|_a, \qquad \delta_W := \|W\|_a,$$

recalling $W = G - \sum_{j=1}^m \alpha_j G_j$, where ξ_n is 1 if n is even and 0 otherwise.

Interestingly, the additional power brought by ξ_n seems to be specific to the multipoint perturbation setting, and improves by one order the obtained error. Therefore, somehow, there is a gain in the convergence order that seems to be for free. Numerically, this additional convergence order can be seen, and shows an important gain of the multipoint perturbation theory compared to the standard perturbation theory. The computational cost of computing the approximation γ_n is just proportional to the number of considered operators G_i, that is m compared to the standard perturbation theory cost.

To conclude, perturbation theory is a powerful tool to provide approximations of density matrices related to operators that are close to one or several well-known operators for which the solutions are explicitly known or can be efficiently and very accurately computed. The multipoint perturbation is specifically tailored to leverage the case when several close operators are known, which is exactly the setting in molecular dynamics simulations. This will be illustrated in Section 4.2.

2.3.3 The Feschbach–Schur map

We now turn to another approximation technique which relies on separating the Hamiltonian into several parts. Unlike in Section 2.3.1, where the Hamiltonian is expressed as a sum of two operators, the Feshbach-–Schur formalism introduces a decomposition of the underlying space into two complementary subspaces. This allows the original problem defined on the full space to be reformulated as a nonlinear eigenvalue problem restricted to a lower-dimensional subspace.

This construction is known as the Schur complement, a standard tool in finite-dimensional linear algebra. When solving a linear system, one can partition the system into two components; by exactly resolving one part, the Schur complement naturally arises. This reduces the original system to a smaller one defined on one of the two complementary subspaces. While the situation is more involved for eigenvalue problems, a similar approach can be employed. Extending this approach to infinite-dimensional spaces introduces additional challenges, primarily due to the need for carefully chosen functional spaces. However, the fundamental structure of the technique remains the same.

For eigenvalue problems, this method originated in works of I. Schur on the Dirichlet problem in planar domains [133] and H. Feshbach, on resonances in nuclear physics [59], and was then independently developed in numerical analysis, computational quantum chemistry and mathematical physics, see [70, 73] with the original techniques called variously the Feshbach projection and Schur complement methods. The error introduced by the dimensionality reduction can be quantified, even in the presence of fairly irregular potentials. As in the contributions related to *a posteriori* error estimation for plane wave discretizations, this is done by splitting the space between low and high frequencies in Fourier space, and choosing as finite dimensional space the low frequency space. Using this splitting, we map the original problem to obtain a reduced, finite dimensional one.

In this section we start by explaining how the technique works before moving to the specific contributions related to this [A21, A22]. We consider a periodic lattice $\mathcal{R}$ and a domain $\Omega \subset \mathbb{R}^d$, and we denote by X_M the subspace of $L^2(\Omega)$ spanned by the eigenfunctions of $-\Delta$ on $\mathcal{R}$, with eigenvalues smaller than ρ_M, as

$$\mathsf{X}_M = \left\{ \sum_{k \in \mathcal{R}^*, |k| \leq M} \hat{u}_k e_k(x) \;\middle|\; \hat{u}_k^* = \hat{u}_{-k},\ \hat{u}_k \in \mathbb{C} \right\}, \tag{2.3.6}$$

where $e_k(x) = |\Omega|^{-\frac{1}{2}} e^{ik\cdot x}$ (see (1.2.9)), $\mathcal{R}^* = \frac{2\pi}{L}\mathbb{Z}^d$, and $\rho_M := \left(\frac{2\pi M}{L}\right)^2$.

Let Π_M be the $L^2(\Omega)$-orthogonal projection onto X_M and $\Pi_M^\perp := \mathrm{id} - \Pi_M$. We consider the Galerkin approximation of the linear Hamiltonian $H := -\Delta + V$,

$$H_M := \Pi_M(-\Delta + V)\Pi_M.$$

Let φ denote an eigenfunction of the Hamiltonian H, we introduce the projections $\varphi_M = \Pi_M\varphi$ and $\varphi_M^\perp = \Pi_M^\perp\varphi$ and project the exact eigenvalue problem

$$H\varphi = \lambda\varphi \tag{2.3.7}$$

onto the subspace X_M and its complement $\mathsf{X}_M^\perp$ to obtain

$$\Pi_M(H_M - \lambda)\varphi_M = -\Pi_M V\varphi_M^\perp, \tag{2.3.8}$$

$$\Pi_M^\perp(H_M^\perp - \lambda)\varphi_M^\perp = -\Pi_M^\perp V\varphi_M, \tag{2.3.9}$$

where $H_M^\perp := \Pi_M^\perp H\Pi_M^\perp$. This system is just the block decomposition of the eigenvalue problem (2.3.7) on X_M and $\mathsf{X}_M^\perp$. Thus for $\lambda < \kappa_M$, where κ_M is such that $H_M^\perp \geq \kappa_M$ on $\mathrm{Ran}\,\Pi_M^\perp$, the operator $H_M^\perp - \lambda$ is invertible and we can solve (2.3.9) for $\varphi_M^\perp$ and thus $\varphi_M^\perp = -(H_M^\perp - \lambda)^{-1}\Pi_M^\perp V\varphi_M$. Substituting the result into (2.3.8), we obtain the non-linear eigenvalue problem

$$\big(H_M + U_M(\lambda)\big)\varphi_M = \lambda\varphi_M, \tag{2.3.10}$$

where we introduced the *effective interaction* $U_M(\lambda) : \mathsf{X}_M \to \mathsf{X}_M$, or a Schur complement,

$$U_M(\lambda) := -\Pi_M V\Pi_M^\perp(H_M^\perp - \lambda)^{-1}\Pi_M^\perp V\Pi_M. \tag{2.3.11}$$

Having shown that the Schur complement is well-defined for $\lambda < \kappa_M$, we can, as in the standard perturbation setting, use a Neumann expansion of the resolvent $(H_M^\perp - \lambda)^{-1}_{|\mathrm{Ran}\Pi_M^\perp} = (-\Delta + V_M^\perp - \lambda)^{-1}_{|\mathrm{Ran}\Pi_M^\perp}$ in (2.3.11) in the formal Neumann series in $V_M^\perp := \Pi_M^\perp V\Pi_M^\perp$.

In our contributions, we then truncate this series at $K \in \mathbb{N}$ and replace the projections $\Pi_M^\perp = \mathrm{id} - \Pi_M$ by $\Pi_M^N := \Pi_N - \Pi_M$, with $N > M$. Introducing the notation

$$\mathsf{G}_M^N(\lambda) := (-\Delta - \lambda)|^{-1}_{\mathrm{Ran}\,\Pi_M^N}, \tag{2.3.12}$$

and $V_M^N := \Pi_M^N V\Pi_M^N$, we obtain the following truncated effective interaction

$$U_\sigma(\lambda) := -\Pi_M V\Pi_M^N R_\sigma(\lambda)\Pi_M^N V\Pi_M, \tag{2.3.13}$$

where $\sigma := (N, M, K)$ and $R_\sigma(\lambda) := \sum_{k=0}^K (-1)^k \Big[\mathsf{G}_M^N(\lambda)V_M^N\Big]^k \mathsf{G}_M^N(\lambda)$. Since all the operators involved in (2.3.13) are finite matrices, this family is well-defined and computable. Now, we define $\mathcal{H}_\sigma(\lambda) := H_M + U_\sigma(\lambda)$ on X_M and consider the eigenvalue problem: find an eigenvalue $\lambda_{\sigma i}$ and the corresponding eigenfunctions $\varphi_{\sigma i} \in \mathsf{X}_M$ such that

$$\mathcal{H}_\sigma(\lambda_{\sigma i})\varphi_{\sigma i} = \lambda_{\sigma i}\varphi_{\sigma i}. \tag{2.3.14}$$

We need one additional definition which is the 'lifting' operator defined as

$$Q_\sigma(\lambda) := \mathrm{id} - R_\sigma(\lambda)\Pi_M^N V\Pi_M. \tag{2.3.15}$$

Our main results are *a priori* error estimates for the approximation error of the eigenvalue and eigenfunction. Indeed we quantify the error introduced due to the discretization parameters $\sigma = (N, M, K)$, see [A21, Theorem 1].

Theorem 2.3.3. *Let $\lambda_\star$ be an isolated eigenvalue of H of finite multiplicity m, with eigenfunctions φ_i, and let δ_0 denote the gap between $\lambda_\star$ and the rest of the spectrum of H.*

Then, there exists $\alpha > 0$ and $M_0 \in \mathbb{N}$ such that for $N \geq M \geq M_0$, problem (2.3.14) has m solutions $(\varphi_{\sigma i}, \lambda_{\sigma i}) \in \mathsf{X}_M \times [\lambda_\star - \frac{\delta_0}{2}, \lambda_\star + \frac{\delta_0}{2}]$ approximating $(\varphi_i, \lambda_\star)$ in the following sense:

$$|\lambda_- \lambda_{\sigma i}| \lesssim (\lambda_\star + \alpha) \frac{\|V\|_r^2}{\alpha^r} \varepsilon(\sigma, r, V), \tag{2.3.16}$$

$$\|\varphi_i - Q_\sigma(\lambda_{\sigma i})\varphi_{\sigma i}\| \lesssim \|V\|_r \left[1 + \frac{\lambda_\circ}{\delta_0} \frac{\|V\|_r}{\alpha^r}\right] \varepsilon(\sigma, r, V), \tag{2.3.17}$$

where $\lambda_\circ = \lambda_+ \delta_0 + \alpha$ and

$$\varepsilon(\sigma, r, V) := \rho_N^{-r} + \rho_M^{-r} \left[4\rho_M^{-r} \|V\|_r\right]^{K+1}.$$

In the considered setting, ε is equivalent to

$$\varepsilon(\sigma, r, V) \approx N^{-2r} + M^{-2r} \left[4 \left(\tfrac{L}{2\pi}\right)^{2r} M^{-2r} \|V\|_r\right]^{K+1},$$

where the equivalence constants do not depend on the parameters $\sigma = (N, M, K), r, \alpha, V$. This indeed provides an *a priori* error bound of both the eigenvalue and the eigenvector error. Naturally convergence of the eigenvalues and the eigenfunctions can be achieved by taking the limit $K, N \to \infty$ for fixed $M \geq M_0$. But in practice, the idea is to set N large enough so that the error is dominated by the error introduced in $K < +\infty$.

Further, note that the eigenvalue and eigenvector errors have the same rate of convergence with respect to K. However, the error in the eigenvector depends on the gap δ_0 while the error in the eigenvalue does not. This is to be compared with the *a posteriori* estimates of [A10] presented in Section 2.2.2 where both the eigenvalue and eigenvector errors depend on the gap, except in the regular enough case where asympotically the error does not depend on the gap (see (2.2.6) for the eigenvalues).

The estimate with respect to N in Theorem 2.3.3 is not sharp in all cases, in particular for sufficiently regular potentials V, but holds for low regularities of the potential where standard *a priori* convergence results of the variational approximation approximation results of the variational approximation result do not hold. Note that in the low-regularity setting *a posteriori* error estimates cannot be derived using the approach presented in Section 2.2.2.

To conclude, perturbation methods are powerful tools in the analysis of eigenvalue problems and can be adapted and extended in various ways. In general, the error estimates that are derived in this context are *a priori* error estimates: they provide with convergence rates, but do not give guaranteed error estimates as derived in the previous sections. However this can be useful anyway to accelerate calculations, such as by providing improved solutions at low computational cost. Indeed one possibility is to perform a simulation in a low-dimensional space, and then to improve the solution in a finer basis using perturbative arguments, similarly in spirit to two-grid or multigrid methods used in finite elements simulations [149, 78, 35].

2.4 Perspectives

I would now like to mention some research perspectives linked to *a posteriori* error estimation and efficient calculations.

- **Localized basis sets** General *a posteriori* error estimates remain to be developed for localized basis sets. First, one could extend what has been done in [A16] to clusters of eigenvalues, as well as quantities of interest, and for nonlinear problems. Another

important consideration is how to perform efficient 3D calculations for atomic systems, and how to obtain optimized basis sets. A logical starting point would be to consider basis sets available in the literature, before optimizing them and providing better atomic basis sets, which, moreover, would be compatible with the *a posteriori* analysis.

- **Non-gapped systems** The presented estimates rely on the assumption that there exists a spectral gap between the eigenvalues of interest and the rest of the spectrum. However this assumption does not always hold, especially for metallic systems. In such cases, it may be possible to rely on a further gap in the spectrum, i.e. between higher eigenvalues. Computing a larger number of eigenvalues is indeed what is done in practice to compute the electronic structure of metals. But one could also consider smearing, which consists of adding temperature to the system, in order to introduce a gap, at the price of making some approximation, which would then have to be estimated.

- **Guaranteed bounds for nonconvex functionals** Often when considering nonconvex problems, only local error bounds can be obtained using standard methods, such as the Newton–Kantorovitch theorem. It would be interesting to determine whether global bounds can also be obtained. Although such error bounds seem out of reach for generic functionals, it may be possible to obtain bounds for specific exchange-correlation functionals, such as when using the local density approximation exchange-correlation functional, which is close to be convex.

- **Band-spectrum guarantee and k-point sampling** For materials systems, we have only focused on single $\mathbf{k}$-points calculations. However, the band-spectrum corresponds to the map $\mathbf{k} \mapsto (\lambda_{\boldsymbol{k},i})_i$. In practice, the Brillouin zone is sampled at a few points and the band spectrum is extrapolated from the calculation at these few points. Quantities of interest are calculated using quadrature rules. It would be useful to provide an *a posteriori* analysis in this context, for which an *a priori* estimation is available in [37]. We could therefore obtain a fully guaranteed band spectrum, and from there obtain guaranteed quantities of interest such as the density of states. With such bounds in hand, we could develop adaptive $\mathbf{k}$-point quadrature rules, in order to obtain optimized grids at low computational cost. In this direction, see e.g. [22].

- **Efficient calculations** Another useful outcome of the error bounds is to allow for parameter optimization in order to reach a target accuracy at a minimal computational cost. For this, practical optimized strategies remain to be developed. Indeed as mentioned in [147], it seems that natural strategies do not always lead to the best compromise. A second step would be to incorporate such strategies in quantum chemsitry codes to maximize their use, starting with freely available codes, such as the plane-wave DFT code DFTK.jl [79], before moving to commercial codes for which accessing the functions is more difficult.

- **Extensions of multipoint perturbation** So far the multipoint perturbation has been proposed for linear eigenvalue problems, and single eigenvalues. Extending this work to multiple eigenvalues should not be too difficult, as long as there is a gap between the considered eigenvalues and the remainder of the spectrum. Extending this work to nonlinear eigenvalue problems and especially electronic structure equations may be more involved but would be very interesting. It may help theoretically understanding how to optimize the approximation of a density matrix with nearby points.

Chapter 3

Optimal transport distances adapted to electronic structure calculations

Originally introduced by Monge in [109] to study the problem of efficiently moving a pile of sand to cover a sinkhole, optimal transport has since been widely developed, and found applications across a broad range of fields, from imaging [123, 49], data science [139, 42], model order reduction techniques [55], to tomographic reconstruction [3]. Applications of optimal transport to electronic structure calculations have been so far rather limited. A few works are concerned with strongly correlated systems where the considered partial differential equation can be recast as a multimarginal optimal transport problem with Coulomb cost instead of the usual quadratic cost [46]. Different numerical methods have been developed for this problem [23, 5, 64].

My research takes a different angle. I am interested in exploiting a very nice feature of Wasserstein barycenters: they provide a way to interpolate between probability measures. These interpolations seem well suited for electronic structure calculations, since electronic densities are probability measures that, moreover, behave smoothly when the nuclei coordinates $\mathbf{R}$ change. There are, however, at least two limitations for the Wasserstein barycenters to be used in electronic structure calculations. First the computational cost of standard Wasserstein barycenters is high in general, and scales badly with the dimension of the physical space, even using regularization techniques such as the Sinkhorn algorithm [117]. Second the electronic density satisfies some structural properties (e.g. symmetry, regularity) that Wasserstein barycenters do not preserve in general. My main goal in this direction is to propose modified Wasserstein distances which are (i) numerically efficient to compute – meaning that the computation complexity of the barycenter should be independent of the physical space dimension d, and (ii) respect given constraints. These distances will be used in Chapter 4 to build efficient nonlinear reduced order models as the one presented in [S1].

3.1 Wasserstein distance and barycenters

We first provide with the necessary tools related to optimal transport, such as Wasserstein metric and barycenters, see for instance [144, 130, 117, 63] for references. Let $\Omega \subset \mathbb{R}^d$ for $d \in \mathbb{N}$ be a Borelian set. We denote by $\mathcal{P}(\Omega)$ the set of probability measures on Ω.

3.1.1 Wasserstein distance

The 2-Wasserstein distance is defined on the set of probability measures with finite second-order moments denoted by $\mathcal{P}_2(\Omega)$ as

$$W_2(\mu_0,\mu_1) := \inf_{\pi\in\Pi(\mu_0,\mu_1)} \left(\int_{\Omega\times\Omega} \|x-y\|^2 \, d\pi(x,y)\right)^{1/2}, \tag{3.1.1}$$

where $\Pi(\mu_0,\mu_1)$ denotes the set of measures on $\Omega\times\Omega$ with marginals μ_0 and μ_1, also called the set of transport plans between μ_0 and μ_1. The space $\mathcal{P}_2(\Omega)$ endowed with the distance W_2 is a metric space, usually called L^2-Wasserstein space (see [144] for more details). From [130, Theorem 1.17], there exists a unique optimal transport plan solution of the minimization problem (3.1.1) denoted by π in the sequel, provided that μ_0 is absolutely continuous with respect to the Lebesgue measure. Also, the optimal transport plan π has the following form

$$\pi = (\mathrm{Id}, T)\#\mu_0,$$

where $T:\Omega\to\Omega$ is an application called the optimal transport map and satisfying $T\#\mu_0 = \mu_1$. Here, we denote by $T\#\mu$ the push-forward measure of a measure μ on Ω by a map $T:\Omega\to\Omega$, that is the measure ν on Ω such that $\forall A\subset\Omega$, $T\#\mu(A) = \mu(T^{-1}(A))$. If μ_0 admits a density the optimal transport plan takes the form

$$\forall (x,y)\in\Omega^2, \quad \pi(x,y) = \delta_{y=T(x)}\mu_0(x).$$

The path $(\mu_t)_{t\in[0,1]}$ given by

$$\forall t\in[0,1], \quad \mu_t = P_t\#\pi = ((1-t)\mathrm{Id} + tT)\#\mu_0, \quad \forall (x,y)\in\Omega^2,\ P_t(x,y) := (1-t)x + ty,$$

defines a constant speed geodesic in $\mathcal{P}_2(\Omega)$ between μ_0 and μ_1. The path $(\mu_t)_{t\in[0,1]}$ is called the McCahn interpolation between μ_0 and μ_1 [106]. It can be shown that for all $t\in[0,1]$, μ_t can be equivalently expressed as the unique solution to the following minimization problem

$$\mu_t := \operatorname*{argmin}_{\mu\in\mathcal{P}_2(\Omega)} (1-t)W_2^2(\mu,\mu_0) + tW_2^2(\mu,\mu_1). \tag{3.1.2}$$

3.1.2 Wasserstein barycenters

We next introduce the notion of barycenters in the Wasserstein space [4] which can be seen as an extension of the McCahn interpolation to a family of more than two measures. Let $n\in\mathbb{N}^*$ and let

$$\Lambda_n := \left\{ \boldsymbol{t} := (t_1,\cdots,t_n)\in[0,1]^n, \quad \sum_{i=1}^n t_i = 1 \right\}$$

be the probability simplex of dimension $n-1$. For any family of probability measures $\boldsymbol{\mu} = (\mu_i)_{1\le i\le n}\in(\mathcal{P}_2(\Omega))^n$ and barycentric weights $\boldsymbol{t} = (t_i)_{1\le i\le n}\in\Lambda_n$, if one of the measures μ_i has a density, there exists a unique minimizer to the problem

$$\inf_{\overline{\mu}\in\mathcal{P}_2(\Omega)} \sum_{i=1}^n t_i W_2(\overline{\mu},\mu_i)^2, \tag{3.1.3}$$

which is the barycenter of the family of measures $\boldsymbol{\mu}$ with barycentric weights $\boldsymbol{t}$. The solutions of the barycenter problem are, moreover, related to the solutions of the following multi-marginal optimal transport problem [65]

$$W_2(\mu_1,\ldots,\mu_n)^2 := \inf_{\pi\in\Pi(\mu_1,\ldots,\mu_n)} \int_{\Omega^n} \frac{1}{2}\sum_{i,j=1}^n t_i t_j \|x_i - x_j\|^2 \, d\pi(x_1,\ldots,x_n), \tag{3.1.4}$$

where $\Pi(\mu_1, \dots, \mu_n)$ denotes the set of probability measures on Ω^n having $\mu_1, \dots, \mu_n$ as marginals. In particular, if Ω is a convex set and if (3.1.4) has a unique solution π^*, there exists a unique solution ρ^* to (3.1.3) given by $\rho^* = B\#\pi^*$, with $B : \Omega^n \to \Omega$ defined by $B(x_1, \dots, x_n) := \sum_{i=1}^n t_i x_i$ for all $(x_1, \dots, x_n) \in \Omega^n$, and the infima of (3.1.3) and (3.1.4) are equal. We denote the Wasserstein barycenter by

$$\mathrm{Bar}^{\boldsymbol{t}}_{W_2}(\boldsymbol{\mu}).$$

In general, computing the Wasserstein distance or Wasserstein barycenter between a family of measures is numerically demanding, especially when the number of measures is even moderately large. Indeed solving (3.1.4) with n measures discretized on grids with N points per dimension d requires to solve a linear programming problem of size n^{Nd} which becomes quickly unfeasible in practice. As a result, the computational cost may render interpolation between electronic densities unfeasible. Fortunately, there are a few cases in which the computation of Wasserstein distance and barycenters is explicit or easy to compute. We now present the case of one-dimensional probability measures, and the case of measures based on translations and dilations of a reference measure (e.g. a Gaussian).

3.1.3 One-dimensional case

For one-dimensional distributions, we denote the cumulative distribution function (cdf) of an element $u \in \mathcal{P}_2(\mathbb{R})$ defined by

$$\mathrm{cdf}_u : x \in \mathbb{R} \longmapsto \int_{-\infty}^{x} \mathrm{d}[u],$$

and the inverse cumulative distribution function (icdf) defined as the generalized inverse of the cdf

$$\mathrm{icdf}_u : s \in [0,1] \longmapsto \mathrm{cdf}_u^{-1} := \inf\{x \in \mathbb{R}, \mathrm{cdf}_u(x) > s\}.$$

Then, for any $(u, v) \in \mathcal{P}_2(\mathbb{R})^2$, there holds

$$W_2(u, v) = \|\mathrm{icdf}_u - \mathrm{icdf}_v\|_{L^2([0,1])}, \tag{3.1.5}$$

and for any set of barycentric weights $\boldsymbol{t} := (t_1, \dots, t_n)$ and $\boldsymbol{u} := (u_1, \dots, u_n)$, the icdf of the barycenter $\mathrm{Bar}^{\boldsymbol{t}}_{W_2}(\boldsymbol{u})$ satisfies

$$\mathrm{icdf}_{\mathrm{Bar}^{\boldsymbol{t}}_{W_2}(\boldsymbol{u})} = \sum_{i=1}^{n} t_i \mathrm{icdf}_{u_i}.$$

Wasserstein distance and barycenters are therefore easy to compute in this case. After the nonlinear transformation of considering the icdf, the Wasserstein distance can be computed as a standard L^2-norm. Unfortunately this property is specific to the one-dimensional case.

3.1.4 Location-scatter distributions

In dimension $d \in \mathbb{N}$, we consider the set of location-scatter distributions consisting of a reference measure that is translated and dilated (see [63]) that are called location-scatter measures [6]. E.g. taking a Gaussian measure as a reference, we obtain the set of Gaussian distributions. These distributions are characterized by a center $m \in \mathbb{R}^d$ and a covariance matrix $\Sigma \in \mathcal{S}_{+,*}$, with $\mathcal{S}_{+,*}$ the set of symmetric positive definite matrices of $\mathbb{R}^{d \times d}$.

For example, for any $m \in \mathbb{R}^d$ and $\Sigma \in \mathcal{S}_{+,*}$, the (normalized) Gaussian distribution $g_{m,\Sigma} \in \mathcal{P}_2(\mathbb{R}^d)$ is defined by

$$g_{m,\Sigma}(\,dx) = G_{m,\Sigma}(x)\,dx,$$

where

$$\forall x \in \mathbb{R}^d, \quad G_{m,\Sigma}(x) := \frac{1}{(2\pi)^{d/2}\sqrt{\det \Sigma}} \exp\left((x-m)^T \Sigma^{-1}(x-m)\right). \tag{3.1.6}$$

For any $m_0, m_1 \in \mathbb{R}^d$ and $\Sigma_0, \Sigma_1 \in \mathcal{S}_{+,*}$, denoting by $g_0 := g_{m_0,\Sigma_0}$ and $g_1 := g_{m_1,\Sigma_1}$, the Wasserstein distance is given by

$$W_2^2(g_0, g_1) = \|m_0 - m_1\|^2 + \mathrm{Tr}\left(\Sigma_0 + \Sigma_1 - 2\left(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2}\right)^{1/2}\right). \tag{3.1.7}$$

Moreover the Wasserstein barycenter between n Gaussian measures is also a Gaussian measure. Precisely, for $\boldsymbol{t} = (t_i)_{1\le i\le n} \in \Lambda_n$, the unique minimizer of (3.1.3) is given by

$$g_{\boldsymbol{t}}(\,dx) = G_{m_{\boldsymbol{t}},\Sigma_{\boldsymbol{t}}}(x)\,dx, \tag{3.1.8}$$

where $m_{\boldsymbol{t}}$ and $\Sigma_{\boldsymbol{t}}$ satisfy

$$m_{\boldsymbol{t}} = \sum_{i=1}^n t_i m_i, \quad \Sigma_{\boldsymbol{t}} = \sum_{i=1}^n t_i \left(\Sigma_{\boldsymbol{t}}^{1/2}\Sigma_i\Sigma_{\boldsymbol{t}}^{1/2}\right)^{1/2}. \tag{3.1.9}$$

This means that the mean of the barycenter is the (euclidean) barycenter of the means, and the covariance matrix satisfies a fixed-point equation. Insights behind this formula can be found in [4, 6]. In practice, this equation can easily be solved with an iterative algorithm. The formulas are identical for other reference measures, only (3.1.6) which contains the reference measure has to be adapted.

This is another important case where explicit formulas are available for both the Wasserstein distance and its barycenters. Interestingly, the computational cost of determining the barycenters does not scale poorly with the physical space dimension d.

3.2 Modified Wasserstein distances using mixtures of location-scatter measures

Location-scatter measures are particularly appealing due to their explicit formulas for the Wasserstein distance and barycenters; however, they do not cover a sufficiently large space for most practical applications. A natural extension consists in considering mixtures of location-scatter measures. Indeed any regular-enough measure can be well approximated as a mixture of such measures. The reference measure is called below an atom. In the literature, the chosen atoms are often Gaussians. For example, a modified Wasserstein distance has been proposed for Gaussian mixtures in [41, 49]. For electronic structure problems other atoms may be better suited. We therefore extended the Gaussian mixture distance to a larger setting of general location-scatter measures in [S2], and we summarize here the main findings: sufficient conditions on the densities constituting the mixtures and the definition of the modified Wasserstein distance on mixtures to define a metric and geodesic space.

3.2.1 Mixture distance and barycenters

We start by providing a generic definition of mixtures.

Definition 3.2.1 (A-mixture)**.** *Let A be a subset of $\mathcal{P}_2(\Omega)$, called dictionary of atoms hereafter. We denote by $\mathcal{M}(A)$ the set of finite mixtures of atoms A, i.e. the set of probability measures μ of $\mathcal{P}(\Omega)$ such that there exists $K \in \mathbb{N}^*$, $\boldsymbol{a} := (a^1, \dots, a^K) \in A^K$ and $\boldsymbol{\lambda} := (\lambda^1, \dots, \lambda^K) \in \Lambda_K$ such that*

$$\mu = \sum_{k=1}^K \lambda^k a^k.$$

A natural example of dictionary of atoms is the set of non-degenerate Gaussian measures

$$A_g^d := \left\{ g_{m,\Sigma},\ m \in \mathbb{R}^d,\ \Sigma \in \mathcal{S}_{+,*} \right\},$$

which is the set $\mathcal{M}(A_g^d)$ of finite Gaussian mixtures of dimension d. In order to define a distance on the set of mixtures as well as establish many properties on the space of mixtures which we study in the sequel, we introduce a map $\delta : A \times A \to \mathbb{R}_+$, which needs to satisfy the following assumption.

Assumption 3.2.2. *The application $\delta : A \times A \to \mathbb{R}^+$ defines a metric on A and (A, δ) is a geodesic space.*

We then define the following mixture distance, for simplicity as an analog of the W_2 distance. It can easily be extended to p-distances.

Definition 3.2.3 (Mixture distance)**.** *Let $A \subset \mathcal{P}(\Omega)$ be a dictionary of atoms, and let $\delta : A \times A \to \mathbb{R}_+$ be a metric on A. We define the application $\delta_{\mathcal{M}} : \mathcal{M}(A) \times \mathcal{M}(A) \to \mathbb{R}_+$ as follows: for all $\mu_0 = \sum_{j=1}^J \lambda_0^j a_0^j \in \mathcal{M}(A)$ and $\mu_1 = \sum_{k=1}^K \lambda_1^k a_1^k \in \mathcal{M}(A)$,*

$$\delta_{\mathcal{M}}(\mu_0, \mu_1)^2 := \min_{w:=(w_{jk})_{\substack{1\le j\le J,\\ 1\le k\le K}} \in \Pi(\Lambda_0,\Lambda_1)} \sum_{j=1}^J \sum_{k=1}^K w_{jk} \delta^2(a_0^j, a_1^k), \tag{3.2.1}$$

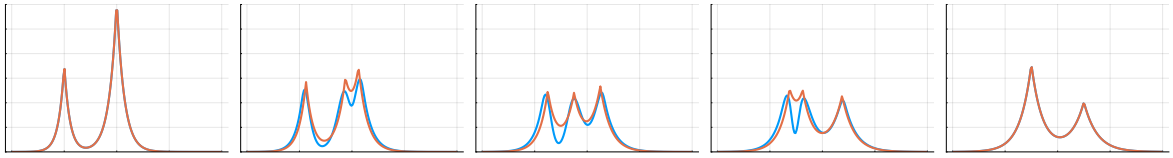
$$\text{with} \quad \Pi(\Lambda_0, \Lambda_1) := \Bigg\{ w := (w_{jk})_{\substack{1\le j\le J,\\ 1\le k\le K}} \in \mathbb{R}_+^{J\times K}, \\ \forall 1 \le j \le J,\ \sum_{k=1}^K w_{jk} = \lambda_0^j, \quad \forall 1 \le k \le K,\ \sum_{j=1}^J w_{jk} = \lambda_1^k \Bigg\}.$$

Roughly speaking, this distance consists solving a discrete transport problem where the ground cost is modified from the Euclidean distance on positions of Dirac measures $\|x_i - x_j\|$ to δ-distances on atoms. This distance is easily computable if the δ-distances can be efficiently computed, which is the case if location-scatter measures are considered for the atoms. When the atom distance is the 2-Wasserstein distance, we denote the mixture distance by MW_2.

In [A16], we have proved that if $\delta : A \times A \to \mathbb{R}^+$ defines a metric, the application $\delta_{\mathcal{M}}$ defines a metric on the set of A-mixtures $\mathcal{M}(A)$, as a simple extension of [49] in the Gaussian case. The computational cost of the distance is very low if the mixtures contain only a few elements, which we expect to be the case for electronic structure calculations.

As shown in [49] this distance can be written as a continuous optimal transport problem similar to the Kantorovitch formulation, where the transport plans are restricted to mixtures of transport plans for the atoms. This property is valid under two main assumptions. The first one is that there indeed exist transport plans between any two elements in the atoms set. This assumption is e.g. satisfied for the cost function $\|x - y\|$ and atoms that are absolutely continuous measures with respect to the Lebesgue measure. The second assumption is that the set of mixtures is identifiable, which means that if two mixtures are equal, then their components are all equal. A classical example of identifiable mixtures is the set of Gaussian mixtures, see e.g. [49, Appendix].

A similar result can be obtained for a mixture multimarginal problem under the same two assumptions, which leads to a definition of mixture Wasserstein barycenters.



(a) $t = 0$ (b) $t = 0.25$ (c) $t = 0.5$ (d) $t = 0.75$ (e) $t = 1$

Figure 3.1: Wasserstein barycenters between two mixtures of Slater-type elliptical distributions for the W_2 metric (blue) and the $W_{2,\mathcal{M}}$ metric (red).

Proposition 3.2.1 (Mixture barycenter)**.** *Under the existence of multi-marginal transport plans, the mixture multi-marginal problem can be considered as*

$$\left(\delta^{\boldsymbol{t}}_{Q,\mathcal{M}}(\mu_1,\dots,\mu_Q)\right)^2 = \inf_{w\in\Pi(\boldsymbol{\lambda}_1,\dots,\boldsymbol{\lambda}_Q)} \sum_{k_1,k_2,\dots,k_Q=1}^{K_1,K_2,\dots,K_Q} w_{k_1,k_2,\dots,k_Q}\left(\delta^{\boldsymbol{t}}_{Q,\mathcal{A}}(a_1^{k_1},a_2^{k_2},\dots,a_Q^{k_Q})\right)^2, \tag{3.2.2}$$

$$\text{with } \Pi(\boldsymbol{\lambda}_1,\dots,\boldsymbol{\lambda}_Q) := \Bigg\{ w\in\mathbb{R}_+^{K_1\times\dots\times K_Q},$$
$$\forall 1\le q\le Q,\ \forall 1\le k_q\le K_q,\ \sum_{k_1,\dots,k_{q-1},k_{q+1},\dots,k_Q=1}^{K_1,\dots,K_{q-1},K_{q+1},\dots,K_Q} w_{k_1,\dots,k_Q} = \lambda_q^{k_q}\Bigg\}.$$

Moreover, any barycenter of $(\mu_1,\dots,\mu_Q)$ with barycentric weights $\boldsymbol{t}$ can be written as

$$\mathrm{Bar}^{\boldsymbol{t}}_{MW_2}(\mu_1,\dots,\mu_Q) = \sum_{\mathbf{k}\in\mathcal{K}} w^*_{\mathbf{k}}(\boldsymbol{t})\ \mathrm{Bar}^{\boldsymbol{t}}_{MW_2}(a_1^{k_1},\dots,a_Q^{k_Q}), \tag{3.2.3}$$

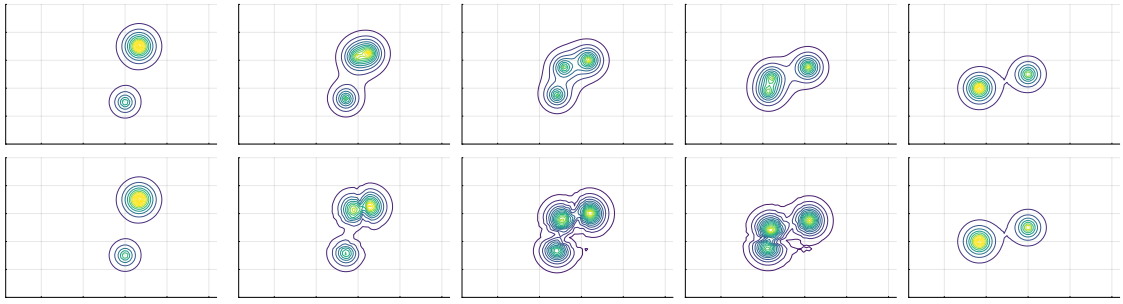
where $\mathrm{Bar}^{\boldsymbol{t}}_{W_2}(a_1^{k_1},\dots,a_Q^{k_Q})$ is the barycenter of $(a_1^{k_1},\dots,a_Q^{k_Q})$ with barycentric weights $\boldsymbol{t}$. Finally, any solution $(w^*_{\mathbf{k}}(\boldsymbol{t}))_{\mathbf{k}\in\mathcal{K}}$ to (3.2.2) contains at most $K_1+K_2+\dots+K_Q-Q+1$ nonzeros components.

This means that modified mixture barycenters can be expressed as convex combinations of barycenters between atoms, which, at least in the case of location-scatter measures with few components, can be very efficiently computed. The mixture barycenter is computed through the resolution of a modified multi-marginal problem, where the ground cost is replaced by the multimarginal cost on atoms. A particularly interesting aspect is that the dimensionality of the problem scales with the number of atoms per mixture, rather than with the ambient spatial dimension.

On top of defining mixtures barycenters based on atoms that are location-scatter measures, the framework is general enough to consider symmetry-adapted atoms when the considered measures satisfy some invariance property. A simple one-dimensional example is to consider even functions. In this case, the atoms can be taken as a sum of two location-scatter measures $a(x)+a(-x)$. This corresponds to a symmetrization over the group action. The atom measure δ and similarly the atom barycenters are in this case computed as simple mixture barycenters.

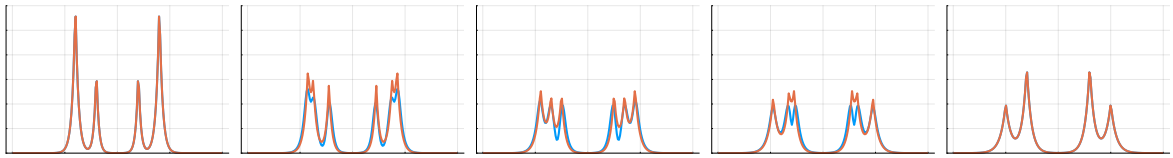
3.2.2 Examples

In [S2], we provided numerical results showing how these mixture Wasserstein barycenters behave. For example, in Figure 3.1 we plot the mixture Wasserstein barycenters and compare



(a) $t = 0$ (b) $t = 0.25$ (c) $t = 0.5$ (d) $t = 0.75$ (e) $t = 1$

Figure 3.2: Contour plots of $W_{2,\mathcal{M}}$ (top) and W_2 (bottom) barycenters between two mixtures of Slater-type elliptic distributions.

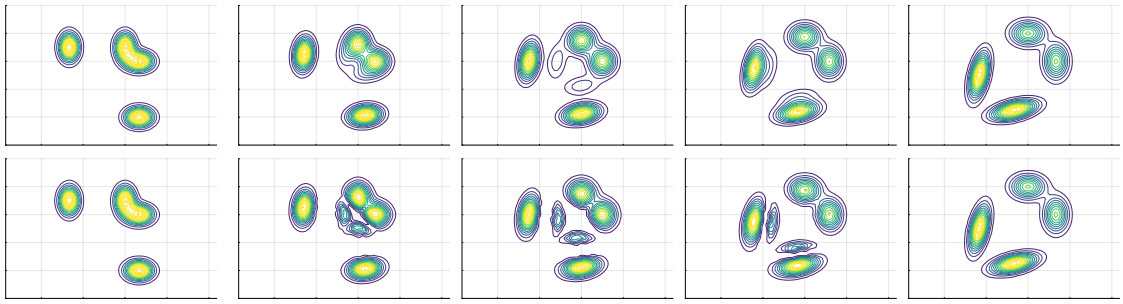


(a) $t = 0$ (b) $t = 0.25$ (c) $t = 0.5$ (d) $t = 0.75$ (e) $t = 1$

Figure 3.3: Wasserstein barycenters between two mixtures of symmetric Slater distributions for the W_2 metric (blue) and the $W_{2,\mathcal{M}}$ metric (red)

them with the true Wasserstein barycenters for Slater-type mixture distributions. The modified barycenter, on top of being way cheaper to compute, keeps the structure of a convex combination of Slater functions. We are therefore now very close to the needed interpolations in the context of the toy problem that will be presented in Chapter 4 (see (4.3.1)). We present in Figure 3.2 a comparison between mixture barycenters and the corresponding true Wasserstein barycenters for a two-dimensional case. Interestingly, the computational cost of the 2d mixture barycenter is almost the same as in the 1d case, whereas it is approximately squared for the true Wasserstein barycenter.

We illustrate the symmetric mixture barycenters in Figures 3.3 and 3.4 respectively in one dimension and two dimensions. In the 2D case, we consider functions satisfying $f(x, y) = f(y, x)$ for any x, y. The chosen definition imposes the modified Wasserstein barycenter to satisfy this symmetry property. This is of importance if one wants e.g. to compute interpolations



(a) $t = 0$ (b) $t = 0.25$ (c) $t = 0.5$ (d) $t = 0.75$ (e) $t = 1$

Figure 3.4: Contour plots of $W_{2,\mathcal{M}}$ (top) and W_2 (bottom) barycenters between two mixtures of symmetric Gaussian distributions.

between pair densities $\rho_{2,\mathbf{R}}$ which satisfy this property.

3.3 Marginal-constrained modified Wasserstein barycenters

Beyond the symmetry $\rho_{2,\mathbf{R}}(x,y) = \rho_{2,\mathbf{R}}(y,x)$ for all x, y, another type of constraint matters for the pair densities: the two marginal of the pair density are the electronic density. Indeed from (1.1.9) there holds

$$\forall x \in \mathbb{R}^3, \quad \int \rho_{2,\mathbf{R}}(x,y)dy = \rho_{\mathbf{R}}(x), \quad \text{and} \quad \forall y \in \mathbb{R}^3, \quad \int \rho_{2,\mathbf{R}}(x,y)dx = \rho_{\mathbf{R}}(y).$$

Unfortunately this is not satisfied in general for Wasserstein barycenters, as shown in [A13]. For applications, we want to define interpolations of pair densities with compatible densities. In [A13], we therefore addressed this problem, first in the case of Gaussian distributions, before defining a modified barycenter satisfying the marginal constraints for mixtures. This opens the way of interpolating between pair densities with exact marginals. The main aspects of this contribution are detailed below.

3.3.1 Marginal-constrained barycenters for Gaussian distributions

For the presentation we assume that the total dimension of the space is $n = n_x + n_y$ for some $n_x, n_y \in \mathbb{N}^*$, so that $\mathbb{R}^n = \mathbb{R}^{n_x} \times \mathbb{R}^{n_y}$. In the case of three-dimensional pair densities, $n_x = n_y = 3$, and so $n = 6$. For a given probability measure $\rho \in \mathcal{P}_2(\mathbb{R}^{n_x} \times \mathbb{R}^{n_y})$, we will denote by $\mathrm{marg}_x(\rho) \in \mathcal{P}_2(\mathbb{R}^{n_x})$ the first marginal of ρ and by $\mathrm{marg}_y(\rho) \in \mathcal{P}_2(\mathbb{R}^{n_y})$ its second marginal. Unlike for pair densities, the two marginal do not have to match in general. More precisely,

$$d\mathrm{marg}_x(\rho)(x) = \int_{y\in\mathbb{R}^{n_y}} d\rho(x,y) \quad \text{and} \quad d\mathrm{marg}_y(\rho)(y) = \int_{x\in\mathbb{R}^{n_x}} d\rho(x,y).$$

If $\rho = \mathcal{N}(\mu, S)$ with $\mu = (\mu_x, \mu_y)$ for some $\mu_x \in \mathbb{R}^{n_x}$ and $\mu_y \in \mathbb{R}^{n_y}$, and $S \in \mathcal{S}_{+,*}^{n_x+n_y}$ written as a block matrix as $S = \begin{pmatrix} S_x & S_{xy} \\ S_{xy}^\mathsf{T} & S_y \end{pmatrix}$ with $S_x \in \mathcal{S}_{+,*}^{n_x}$, $S_y \in \mathcal{S}_{+,*}^{n_y}$ and $S_{xy} \in \mathbb{R}^{n_x \times n_y}$, it holds that

$$\mathrm{marg}_x(\rho) = \mathcal{N}(\mu_x, S_x) \quad \text{and} \quad \mathrm{marg}_y(\rho) = \mathcal{N}(\mu_y, S_y).$$

We therefore consider the optimization problem of finding the distribution $\rho = \mathcal{N}(\mu, S)$ minimizing the Wasserstein distance between a given Gaussian distribution $\rho_{\mathrm{ref}} = \mathcal{N}(\mu_{\mathrm{ref}}, S_{\mathrm{ref}})$ and a Gaussian distribution that satisfies marginal constraints:

$$\mathrm{marg}_x(\rho) = \mathcal{N}(\mu_x, S_x) \quad \text{and} \quad \mathrm{marg}_y(\rho) = \mathcal{N}(\mu_y, S_y).$$

A partial answer is that due to the marginal constraints

$$\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \quad S = \begin{pmatrix} S_x & Z \\ Z^T & S_y \end{pmatrix}, \tag{3.3.1}$$

for some matrix $Z \in \mathbb{R}^{n_x \times n_y}$. Interestingly the solution to the considered optimization problem is unique and can be expressed using the extra definition of the geometric matrix mean [25] between two matrices S and T, defined as the matrix

$$S\#T := S^{1/2}\left(S^{1/2}T^{-1}S^{1/2}\right)^{-1/2}S^{1/2}. \tag{3.3.2}$$

The full result reads as follows.

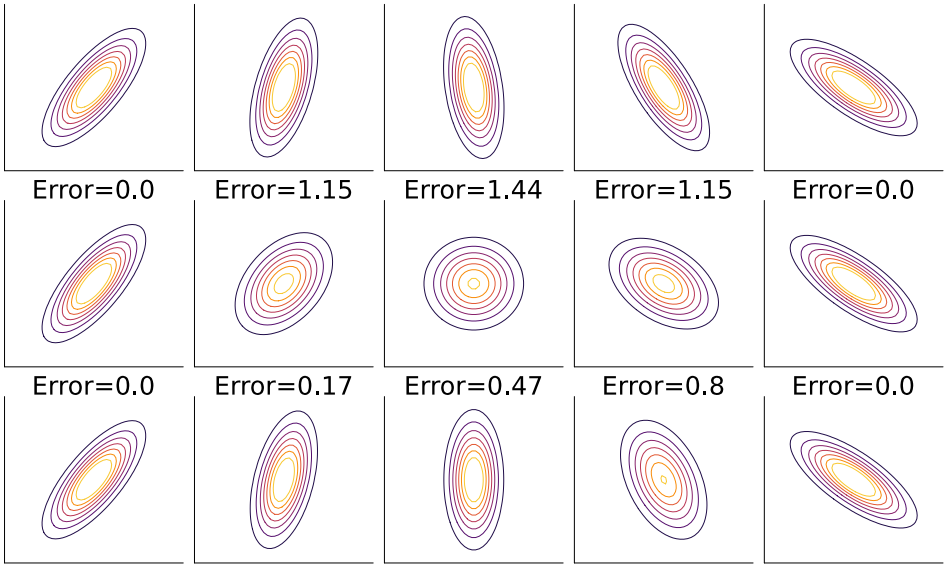


Figure 3.5: Top row: A rotating Gaussian distribution. Middle row: Approximation using a Wasserstein barycenter between the leftmost and rightmost Gaussians with barycentric parameters $0, 0.25, 0.5, 0.75, 1$. Bottom row: Approximation using a marginal-constrained Wasserstein barycenter between the leftmost and rightmost Gaussians with barycentric parameters $0., 0.25, 0.5, 0.75, 1$. The indicated errors correspond to L^2-norms between the considered Gaussian and the corresponding Gaussian on the top row.

Theorem 3.3.1. *Let $n_x, n_y \in \mathbb{N}^*$ and let $S, T \in \mathcal{S}_{+,*}^{n_x+n_y}$ with block decomposition*

$$S = \begin{pmatrix} S_x & S_{xy} \\ S_{xy}^\intercal & S_y \end{pmatrix} \qquad \text{and} \qquad T = \begin{pmatrix} T_x & T_{xy} \\ T_{xy}^\intercal & T_y \end{pmatrix}, \tag{3.3.3}$$

with $S_x, T_x \in \mathcal{S}_{+,*}^{n_x}$, $S_y, T_y \in \mathcal{S}_{+,*}^{n_y}$ and $S_{xy}, T_{xy} \in \mathbb{R}^{n_x \times n_y}$. Let

$$\forall Z \in \mathcal{C}_{T_x,T_y} := \left\{ Z \in \mathbb{R}^{n_x \times n_y},\ \|(\sqrt{T_x})^{-1} Z (\sqrt{T_y})^{-1}\|_2 < 1 \right\}, \quad T(Z) := \begin{pmatrix} T_x & Z \\ Z^\intercal & T_y \end{pmatrix}. \tag{3.3.4}$$

The function F defined as the Bures–Wasserstein metric $\mathcal{W}_2$ between S and $T(Z)$

$$F : \mathcal{C}_{T_x,T_y} \ni Z \longmapsto \mathcal{W}_2(S, T(Z))^2 = \mathrm{Tr}\left(S + T(Z) - 2\sqrt{\sqrt{S} T(Z) \sqrt{S}} \right) \tag{3.3.5}$$

is strictly convex. Moreover, the minimization problem

$$Z_{S,T}^* \in \underset{Z \in \mathcal{C}_{T_x,T_y}}{\mathrm{argmin}}\ \mathcal{W}_2^2(S, T(Z)), \tag{3.3.6}$$

has a unique minimizer which is given by

$$Z_{S,T}^* = ({S_x}^{-1} \# T_x) S_{xy} ({S_y}^{-1} \# T_y). \tag{3.3.7}$$

Combining (3.3.1) and (3.3.7) and choosing as the reference covariance matrix the covariance matrix of a Wasserstein barycenter between given distributions, we compute a modified barycenter that exactly satisfies the marginal constraints. Numerically, this amounts to find

the true Wasserstein barycenter of the $n_x + n_y$-dimensional Gaussians and then postprocess this barycenter to satisfy the marginal constraints.

Such a procedure is useful in at least two cases: (i) when the distributions translate but not at constant speed, and (ii) when the distributions rotate. We provide in Figure 3.5 an example of a rotating Gaussian, and show how using the information of the marginals improves the error compared to a classical Wasserstein barycenter between the leftmost and rightmost distributions.

3.3.2 Marginal-constrained barycenters for Gaussian mixtures

We now extend this notion of marginal-constrained barycenter to the case of Gaussian mixtures. Let us assume that the marginals we wish to impose in the constraints are given as Gaussian mixtures under the following form:

$$\sigma_x = \sum_{k=1}^{K} \alpha_x^k \mathcal{N}(\mu_x^k, S_x^k) \quad \text{and} \quad \sigma_y = \sum_{n=1}^{L} \beta_y^n \mathcal{N}(\mu_y^n, S_y^n),$$

for some $K, L \in \mathbb{N}^*$, $\boldsymbol{\alpha}_x := (\alpha_x^k)_{1\leq k\leq K} \in \Lambda_K$, $\boldsymbol{\beta}_y := (\beta_y^n)_{1\leq n\leq L} \in \Lambda_L$. In addition, for all $1 \leq k \leq K$, $\mu_x^k \in \mathbb{R}^{n_x}$ and $S_x^k \in \mathcal{S}_{+,*}^{n_x}$, and for all $1 \leq n \leq L$, $\mu_y^n \in \mathbb{R}^{n_y}$ and $S_y^n \in \mathcal{S}_{+,*}^{n_y}$.

We first write the modified Wasserstein barycenter ρ as a mixture of Gaussians. We then characterize the set of mixtures having given marginals σ_x, and σ_y. We write down an optimization problem: we minimize the Wasserstein distance between a reference density ρ_{ref} and the set of densities satisfying the marginal constraints. Unfortunately this yields a too complex optimization problem. Therefore we select a subset of the densities satisfying marginal constraints based on previously defined modified individual Gaussian distributions satisfying marginal constraints. Expressing the optimization problem over this smaller set renders its resolution computationally feasible, and numerically requires the resolution of an additional linear programming problem of reasonable size.

Numerically, the results are very satisfactory. We first present results the approximation of solutions of a Fokker–Planck equation. We consider the advection-diffusion equation in two dimensions which reads: for $t \in (0, 1)$ and $\mathbf{x} \in \mathbb{R}^2$, find $\psi \in L^1((0, 1); \mathbb{R}^2)$ satisfying

$$\frac{\partial \psi}{\partial t} = -\nabla \cdot (A\mathbf{x}\psi) + D\Delta\psi.$$

with $D > 0$ and $A = \omega \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, $\omega \in \mathbb{R}$. In the simulations, we took $D = 0.1$ and $\omega = 1$. If the initial condition is a Gaussian distribution $\mathcal{N}(r_0, \Sigma_0)$ that is

$$\psi(\mathbf{x}, 0) = \frac{1}{2\pi\sqrt{\det(\Sigma_0)}} \exp\left(-\frac{1}{2}(\mathbf{x} - r_0)^\top \Sigma_0^{-1}(\mathbf{x} - r_0)\right),$$

then the unique solution of the equation is known and reads at each time t with

$$r(t) = R(t)r_0,$$

$$R(t) = \begin{pmatrix} \cos(\omega t) & -\sin(\omega t) \\ \sin(\omega t) & \cos(\omega t) \end{pmatrix},$$

and

$$\Sigma(t) = R(t)^T \Sigma_0 R(t) + 2DtI.$$

Since the Fokker–Planck equation is linear, the solution of this equation can be analytically obtained as soon as initial conditions are taken as Gaussian mixtures. In the numerical tests

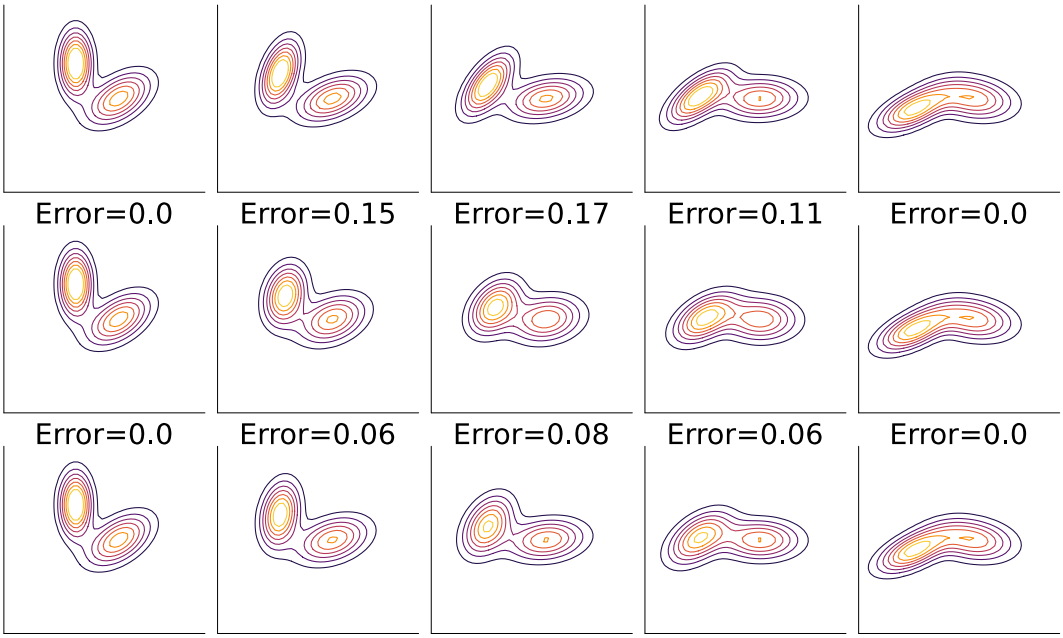


Figure 3.6: Top row: Five snapshots of the time evolution of the solution of a Fokker–Planck equation with an initial solution being a convex combination of two Gaussians. Middle row: Approximation using a Wasserstein barycenter between the leftmost and rightmost solutions with barycentric parameters $0., 0.25, 0.5, 0.75, 1.$. Bottom row: Approximation using a marginal-constrained Wasserstein barycenter between the leftmost and rightmost solutions with barycentric parameters $0., 0.25, 0.5, 0.75, 1.$. The indicated errors correspond to L^2-norms between the considered solution and the corresponding reference solution on the top row.

presented below, we take for initial condition a convex combination of two equally weighted Gaussians: $\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}2 & 1\\1 & 2\end{pmatrix}\right)$ and $\mathcal{N}\left(\begin{pmatrix}-3\\3\end{pmatrix}, \begin{pmatrix}0.5 & 0\\0 & 3\end{pmatrix}\right)$. We compute the solution at times $t = 0, 0.25, 0.5, 0.75, 1$. The corresponding distributions are shown on the top row of Figure 3.6. We then interpolate between the times $t = 0$ and $t = 1$ using two different methods. On the middle row of Figure 3.6, we compute the Wasserstein barycenter between the solutions at $t = 0$ and $t = 1$ with barycentric weights corresponding to time instants equal to $0, 0.25, 0.5, 0.75, 1$. We then plot on the bottom row of Figure 3.6 the barycenter obtained with the same weights but imposing the marginal constraints. We observe that the approximations are more accurate with the latter.

For electronic structure problems, and in particular the approximation of the pair density $\rho_{2,\mathbf{R}}$ for varying $\mathbf{R}$'s, we tested this approach. We first computed solutions to the electronic Schrödinger equation with 3 electrons using a finite-element discretization performed with the Julia package SchrodingerFE.jl [122]. Here an extra fitting step is needed since the pair density is not given as a Gaussian mixture. Therefore we first fit the finite element solutions with 30 Gaussians, and impose the symmetry with respect to the exchange of the variables x and y. The exact pair densities and corresponding Gaussian mixture approximations are respectively given on the first and second row of Figure 3.7. Once again we observe that the approximations given by standard Wasserstein barycenters (third row of Figure 3.7) are cruder than the marginal-constrained Wasserstein barycenters (fourth row). As a side remark, the fit that is done with

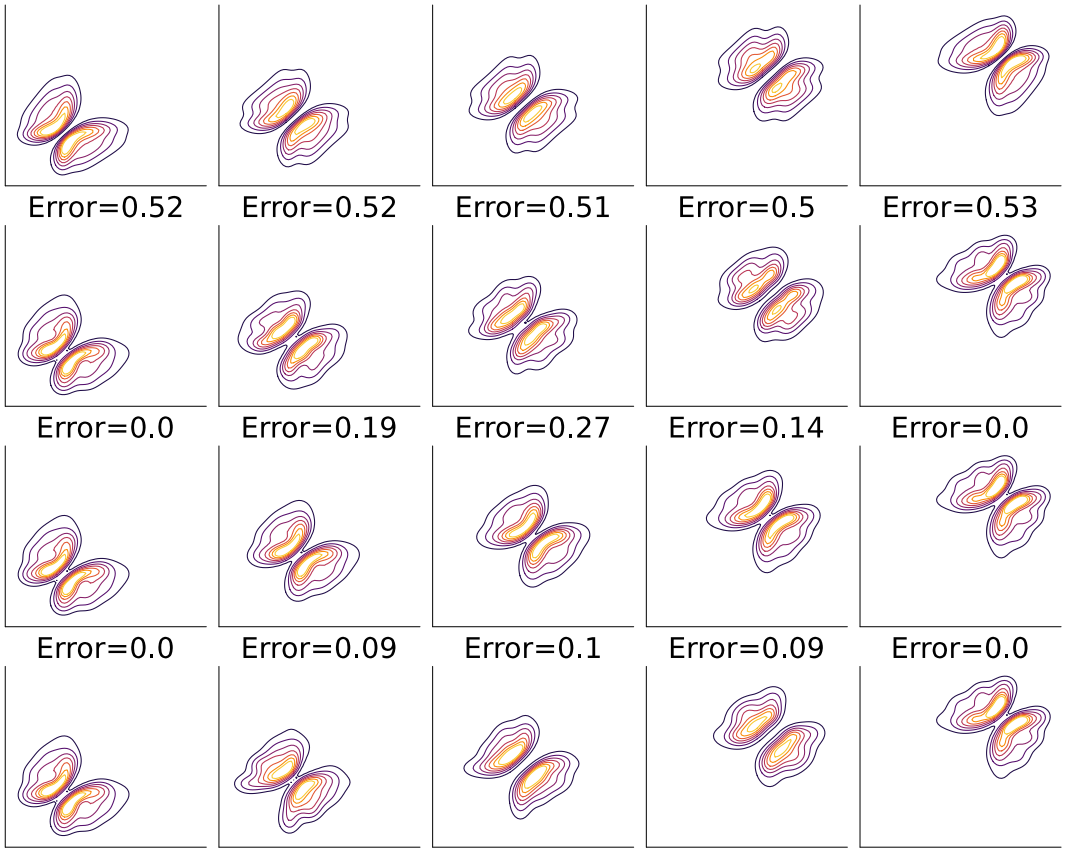


Figure 3.7: Top row: Five snapshots of the time evolution of the solution of a parametrized Schrödinger equation. Second row: Fitting of the solutions with Gaussian mixtures. L^2-error is with respect to first row. Third row: Approximation of the Gaussian mixture fit using a Wasserstein barycenter between the leftmost and rightmost mixture fits with barycentric parameters $0., 0.25, 0.5, 0.75, 1.$. L^2-errors are with respect to second row. Fourth row: Approximation of the Gaussian mixture fit using a marginal-constrained Wasserstein barycenter between the leftmost and rightmost solutions with barycentric parameters $0., 0.25, 0.5, 0.75, 1$. Indicated L^2-errors are with respect to second row.

the standard method in Scikitlearn does not give accurate results.

To conclude it is possible to define modified Wasserstein distances with two appealing properties: they satisfy given constraints (symmetry, marginal, regularity), and the computational cost of both the proposed distances and barycenters remain low, and more importantly independent of the dimension of the physical space. This is extremely encouraging for applications in electronic structure calculations, as will be presented in the Chapter 4 when used in the context of nonlinear reduced order modeling.

3.4 Perspectives

I now mention perspectives related to optimal transport, and the construction of distances tailored for electronic structure calculations.

- **Quantum optimal transport** The natural output of quantum chemistry codes is the density matrix, as indicated in Chapter 1. It would therefore be interesting to provide with interpolation techniques on density matrices relying on a distance designed for them, while keeping similarities to the Wasserstein distance. To do so the recently developed Quantum Optimal Transport (QOT) [68, 45] should provide with answers. However the theory of QOT remains so far limited, and to the best of my knowledge, a metric similar to the Wasserstein metric is still in progress. Moreover at that point barycenters for such metrics are still to be developed, which we intend to do. On top of this, it will be necessary to develop efficient algorithms to compute the barycenters, possibly extending the Sinkhorn algorithm in the classical optimal transport case to the quantum case [117]. Limiting the computational cost of these quantum barycenters could also be done by extending recent works on mixtures of Gaussians [49],[S2] to QOT.

- **Inverse optimal transport** Another perspective would be to optimize the chosen Wasserstein-like metric, similarly to what is proposed in [8]. If done numerically, this would yield a metric adapted to the given data which in our case would come from electronic structure calculations. More generally, it would be interesting to theoretically understand if one can derive a distance directly from the Schrödinger equation or DFT equations, similarly to what is done with gradient flows for other equations.

- **Gaussian mixture fitting** A third perspective is to improve existing algorithms for providing Gaussian mixture fitting, noting that in our framework, the exact probability density is available, while in most of available algorithms, such as the well-known expectation-minimization algorithm, the probability density is only available through samples. Another specificity of electronic structure calculations is that the density may be a linear combination of Gaussians instead of a mixture, i.e. allow for negative coefficients. This is the case when localized Gaussian basis sets are used: some coefficients in the linear combinations for the orbitals may be negative so that the density written as a square of the orbitals contains negative coefficients as well. Simply removing these coefficients does not lead to satisfactory results.

Chapter 4

Nonlinear model order reduction

For a wide range of engineering problems, when a partial differential equation has to be solved many times for different parameter values, one can resort to model order reduction techniques such as the reduced basis method [11, 81]. This approach involves identifying representative snapshots, that is, solutions corresponding to specific parameter values $\mathbf{R}$, to approximate the manifold of solutions. For a new parameter value, the solution is then approximated within the vector space spanned by these selected snapshots. The resolution is expected to be efficient due to the low dimensionality of this so-called reduced space. The accuracy of the approximate solution largely depends on the geometry of the solution manifold, in particular its "flatness," which governs how rapidly the error decreases as the dimension of the reduced space increases. This behavior is formally described by the Kolmogorov width.

In electronic structure calculations, such traditional linear reduced basis methods do not work well for several reasons. In some cases, the considered solutions do not belong to vector spaces, e.g. if one considers density matrices. Second the electronic density is localized around the positions of the nuclei, a phenomenon that is similar to what happens for transport equations [55]. This means that linear combinations of electronic densities (or density matrices) for different parameters $\mathbf{R}$'s are far from looking like an electronic density for different nuclei positions parameters, at least when the nuclei configurations are not very close.

Recently, many works have proposed reduced order models based on different nonlinear transformations. We start this chapter by presenting an overview of linear and recent nonlinear model reduction techniques in Section 4.1. We then present in Section 4.2 a series of contributions [A26, A25, A24] aiming at accelerating *ab initio* molecular dynamics, by proposing improved initial guesses to the self-consistent field algorithm, which reduces the number of iterations needed to reach convergence. The method can be seen as a nonlinear reduced order model, where the non-linearity comes from working on the tangent space of the Grassmann manifold to which the density matrices belong. In Section 4.3 we present a nonlinear reduced order model (see [S1]) for a one-dimensional toy problem in electronic structure, which relies on modified Wasserstein barycenters presented in Chapter 3. In Section 4.4 we present a recent contribution [A18] for the approximation of the density to pair-density map that relies on another nonlinear transformation closely linked to optimal transport, which is known in statistics under the name of copulas. Finally, we present research perspectives in Section 4.5.

4.1 Reduced order modeling

For the sake of the presentation, we first describe the context and goal in reduced order modeling, before reviewing key aspects of linear (also called projection-based) model order reduction, and finally presenting extensions to nonlinear model order reduction.

4.1.1 Context and goal

In reduced order modeling, the main goal is to approximate the manifold of solutions

$$\mathcal{M} := \{\Phi_{\mathbf{R}} \text{ for } \mathbf{R} \in \mathcal{C}\},$$

with $\mathcal{C}$ the set of possible parameters, and $\Phi_{\mathbf{R}}$ the solution to the equations of interest. In the context of quantum chemistry, the set $\mathcal{C}$ is the set of nuclei configurations, and $\Phi_{\mathbf{R}}$ is the electronic structure of the molecule having $\mathbf{R}$ as nuclei positions.

Such a situation arises in many different contexts, such as when exploring or optimizing on some parameter space. In quantum chemistry, simulations performed for many $\mathbf{R}'$s appear in several cases, such as *ab initio* molecular dynamics or in geometry optimization, i.e. where the problem is to find the geometry of the molecule with total lowest energy. Also, it is important to note that the training of machine-learned interatomic potentials presented in Chapter 5 often relies on data coming from electronic structure calculations performed for many $\mathbf{R}$'s.

Here we are interested in understanding the structure of the manifold $\mathcal{M}$, and efficiently approximating elements on this manifold.

4.1.2 Recalling some aspects of linear reduced order modeling

In projection-based reduced order modeling [81], the main idea is to approximate all solutions in $\mathcal{M}$ as linear combinations of basis functions spanning a low-dimensional subspace. The reduced order modeling is done in two steps. In an offline part, where one can afford expensive calculations, a small reduced basis is selected (two methods will be presented below). In the online part, that is when the partial differential equation has to be solved for many parameters $\mathbf{R}$'s, the solution is approximated on the selected space of low dimension, i.e. the problem is projected on the offline-selected reduced basis. The solutions computed in the online phase are therefore elements on this low-dimensional vector space. Hence the linear reduced order modeling only has a chance to work if any element in $\mathcal{M}$ can be well approximated on this low-dimensional vector space. This property can be linked to the notion of Kolmogorov n-width.

Kolmogorov n-width

The Kolmogorov n-width of a manifold $\mathcal{M}$ measures the maximum error made on the best n-dimensional space approximating it. We provide the definition in the context of a Hilbert space $\mathbb{H}$, where the Kolmogorov n-width is defined as the quantity

$$\varepsilon_n(\mathcal{M}, \mathbb{H}) := \inf_{\substack{V_n \subset \mathbb{H} \\ \dim V_n = n}} \sup_{\mathbf{R} \in \mathcal{C}} \|\Psi_{\mathbf{R}} - \mathrm{P}_{V_n} \Psi_{\mathbf{R}}\|, \tag{4.1.1}$$

where P_{V_n} denotes the orthogonal projection on the space V_n. The fast decay of this quantity with respect to n indicates that the projection-based model order reduction has a good chance to work. Moreover, the faster this quantity decays, the better for the reduced order modeling.

In general, evaluating this quantity precisely is very difficult. However, in some cases, one can prove a specific decay of the Kolmogorov n-width. A typical example where the Kolmogorov n-width decays very fast is the case of an elliptic equation

$$A_{\mathbf{R}} \Phi_{\mathbf{R}} = f,$$

where the operator $A_{\mathbf{R}}$ can be decomposed in an affine way as

$$A_{\mathbf{R}} = \sum_{q=1}^{Q} \theta_q(\mathbf{R}) A_q, \quad \text{for some } \theta_q \in \mathbb{R},\ A_q \text{ continuous operators.}$$

In this case the Kolmogorov n-width satisfies for some $c > 0$,

$$\varepsilon_n(\mathcal{M}, \mathbb{H}) \lesssim \exp\Big(-cn^{1/Q}\Big),$$

i.e. decays exponentially with n, with a factor of $1/Q$, where Q is the number of operators in the affine decomposition (see [43, 114] for more details and a proof of this result).

There are cases where the Kolmogorov n-width decays more slowly. A simple example is the one-dimensional transport equation where the parameter is denoted by $\mathbf{R} \in [0, 1]$, and the considered equation is

$$\begin{cases} \partial_t \Psi_{\mathbf{R}}(t, x) + \mathbf{R}\partial_x \Psi_{\mathbf{R}}(t, x) = 0, \quad x \in \mathbb{R},\ t \in \mathbb{R}_+. \\ \Psi_{\mathbf{R}}(0, x) = \mathbf{1}_{[-1,0]} \end{cases}$$

At $t = 1$, the solutions are $\Phi_{\mathbf{R}}(t = 1, x) = \mathbf{1}_{[\mathbf{R}-1,\mathbf{R}]}$, and it can be proven that the Kolmogorov n-width for the manifold $\mathcal{M} := \{\mathbf{1}_{[\mathbf{R}-1,\mathbf{R}]}, \quad \mathbf{R} \in [0, 1]\}$ satisfies for some $c > 0$,

$$\varepsilon_n(\mathcal{M}, L^2(\Omega)) \geq cn^{-1/2},$$

that is exhibits a very slow decay.

Numerical aspects

In practice, one needs to find an appropriate low-dimensional space on which to project the problem in the online part. Often, one first assembles a database of accurately computed solutions for training parameters denoted by $\mathcal{C}_{\text{train}}$. Then there are mainly two options to find the projection space.

The first one is to perform a singular value decomposition of the matrix containing all training solutions, usually called snapshots, i.e. $\Phi_{\mathbf{R}}, \mathbf{R} \in \mathcal{C}_{\text{train}}$ (called POD for proper orthogonal decomposition). The fast decay of the singular values is a good indicator that the projection-based reduced model will work. Moreover, the first singular vectors provide with a good reduced basis. The main drawback of this method is that it requires to construct a potentially large database of accurate solutions.

The second method to construct a reduced space is to perform a greedy algorithm, which selects snapshots one by one in an iterative process. One starts with selecting one parameter $\mathbf{R}_1 \in \mathcal{C}$ and its corresponding solution. Then at each iteration, the snapshot that is worse approximated in the space spanned by the previously selected snapshots $\Phi_{\mathbf{R}_1}, \ldots, \Phi_{\mathbf{R}_{K-1}}$ is selected. An interesting feature of this algorithm is that it does not always require to compute a full database of snapshots. Indeed, if *a posteriori* error estimates are available to estimate the error between a solution in a reduced space and the exact solution, then one can select the solution with the largest error estimate, and only then compute the corresponding reference solution. This means that one only needs to compute accurate solutions for the number of selected solutions in the reduced basis. When the reference solutions are very expensive to compute, this can lead to a huge gain of numerical computations in the offline phase.

4.1.3 Nonlinear model order reduction

Recently, a very active research subject has been the development of nonlinear reduced order methods which are well suited when the Kolmogorov n-width defined in (4.1.1) decays slowly. Unlike in the linear framework, there is currently no standard method available, and the existing approaches are often tailored to specific applications.

In [29], a library-based approximation is presented. The authors consider a reduced order model which is a collection of several linear approximations, so that depending on the parameter

$\mathbf{R}$, different linear spaces can be used. A similar method is to consider tree-based nonlinear reduced modeling [72].

Other contributions rely on different encoding and decoding maps. In [24], a nonlinear decoding map based on compositional polynomial networks is used, while in [82, 99], neural networks are considered. It is also possible to learn the map of coefficients in a large linear reduced basis from the coefficients in a small reduced basis which are easy to compute but do not lead to accurate enough results [10, 44].

Some works have leveraged concepts from optimal transport, which is especially effective in handling translations of solutions, particularly poorly approximated with linear spaces. In [55], Wasserstein barycenters are used in 1D for the approximation of solutions of transport equations. In [87], the approximations are not directly based on optimal transport but incorporate displacement interpolation, which is closely related to it. In the article [S1], the proposed method is based on optimal transport, and more precisely modified 2-Wasserstein distances presented in Chapter 3.

There exist several ways to generalize the notion of Kolmogorov n-width in the nonlinear framework, as presented in [44, Section 2], see also [51]. In this work, the general setting consists of a solution manifold immersed in a Hilbert space $\mathbb{H}$. The adopted framework introduces two continuous maps, an encoding

$$E : \mathbb{H} \to \mathbb{R}^n,$$

and a decoding map

$$D : \mathbb{R}^n \to \mathbb{H},$$

which define a nonlinear reduced order model through the composition of the encoding and decoding maps $D(E(\cdot))$. The parameter $n \in \mathbb{N}$ is the underlying dimension of the reduced model. The maximum error on the manifold $\mathcal{M}$ is given by the quantity

$$\max_{\mathbf{R}\in\mathcal{R}} \|\Phi_{\mathbf{R}} - D(E(\Phi_{\mathbf{R}}))\|,$$

and an extension of the Kolmogorov n-width consists in taking the minimum over possible encoder and decoder maps as

$$\inf_{D,E} \max_{\mathbf{R}\in\mathcal{R}} \|\Phi_{\mathbf{R}} - D(E(\Phi_{\mathbf{R}}))\|.$$

Different notions of widths then arise depending on the constraints and regularity imposed on the maps E and D. For example, if D is linear, we recover the classical notion of Kolmogorov n-width. Then, if E is linear and D can be nonlinear, we obtained the so-called sensing numbers defined as

$$s_n(\mathcal{M}) := \inf_{D,\lambda_1,\dots,\lambda_n} \max_{\Phi\in\mathcal{M}} \|\Phi - D(\lambda_1(\Phi),\dots,\lambda_n(\Phi))\|,$$

where the infimum is taken over all linear functionals $\lambda_1,\dots,\lambda_n \in \mathbb{H}'$ and decoding maps D.

In [55], the authors introduced an extension of the Kolmogorov n-width, where the classical L^2 Kolmogorov n-width is not directly applied to the solutions Φ, but to their inverse cumulative distribution functions (icdf). In the context of that work, the solutions Φ are probability measures, and evaluating the L^2 errors on the icdf's corresponds to computing Wasserstein distances (3.1.5).

In [S1], we extended this notion further to any metric space. Without an underlying Hilbert space, compared to (4.1.1), we need to (i) replace the norm by a distance, (ii) find an alternative for the projection, and (iii) find a replacement for the linear combination. To do so, we strongly

use the notion of barycenter, which is defined for a metric space $\mathbb{M}$, with distance d, for convex parameters $\boldsymbol{t} = (t_1, \ldots, t_n)$, i.e. positive with $\sum_{i=1}^n t_i = 1$, and elements $\Phi_{\mathbf{R}_1}, \ldots, \Phi_{\mathbf{R}_n} \in \mathbb{M}$, as

$$\mathrm{bar}(\boldsymbol{t}; \Phi_{\mathbf{R}_1}, \ldots, \Phi_{\mathbf{R}_n}) := \mathrm{argmin}_{u \in \mathbb{M}} \quad \sum_{i=1}^n t_i d(u, \Phi_{\mathbf{R}_i})^2. \tag{4.1.2}$$

We define the metric Kolmogorov n-width as

$$\varepsilon_n(\mathcal{M}, \mathbb{M}) := \inf_{\Phi_{\mathbf{R}_1}, \ldots, \Phi_{\mathbf{R}_n} \in \mathbb{M}} \sup_{\mathbf{R} \in \mathcal{R}} \inf_{\boldsymbol{t}} d\left(\Phi_{\mathbf{R}}, \mathrm{bar}(\boldsymbol{t}; \Phi_{\mathbf{R}_1}, \ldots, \Phi_{\mathbf{R}_n})\right).$$

Compared to the general definition involving both encoding and decoding maps, this approach corresponds to fixing the decoding map as the Wasserstein barycenter, while imposing no constraint on the encoding map. Theoretically, it is shown in [55] that the metric Kolmogorov n-width decreases very fast for simple 1D transport equations. In [S1], we showed an improvement of the metric Kolmogorov n-widths over the linear width for a toy problem in electronic structure calculations.

4.2 Extrapolation on the Grassmann manifold

In this section we are interested in approximating the manifold of solutions to the Kohn–Sham equations for varying nuclei configurations $\mathbf{R}$, in the context of Born–Oppenheimer Molecular Dynamics (BOMD) (or *ab initio* molecular dynamics). This means that the nuclei configurations $\mathbf{R}$ follow the dynamics of the considered molecule. We construct accurate approximations of the solutions for new configurations $\mathbf{R}$. However unlike standard reduced order models where these approximate solutions are used directly, we use these approximations as initial guesses for the SCF algorithm presented in (1.2.13). Thus we develop a dedicated extrapolation method. We start by detailing the context of *ab initio* molecular dynamics simulations in Section 4.2.1 and we present initial guesses from the literature before detailing the specific contributions in Section 4.2.2.

4.2.1 Born–Oppenheimer molecular dynamics simulations

In BOMD simulations, a time-stepping scheme is employed to simulate the motion of the nuclei within the molecular system under consideration. At each time step t_n, the Kohn–Sham equations are solved for the current nuclei configuration $\mathbf{R}_n$. The density matrix solution to the Kohn–Sham equations denoted by $D_n := D_{\mathbf{R}_n}$ is therefore computed by solving equations (1.2.12). Then from the density matrix, the forces applied to the nuclei are computed, from which are deduced the new positions of the nuclei via the second law of Newton. Then the Kohn–Sham equations are solved for the next iteration of the new nuclei configurations $\mathbf{R}_{n+1}$.

As presented in Section 1.1.1, the Kohn–Sham equations (1.1.4) are solved using an iterative SCF algorithm (1.2.13). Therefore, the overall cost of the SCF algorithm is proportional to the number of iterations needed to reach convergence, often characterized by a chosen norm between two iterates being smaller than a given tolerance. Since the number of iterations is directly related to how close the initial guess is to the converged solution, providing accurate initial guesses is crucial for reducing the overall computational cost.

In the context of molecular dynamics simulations, an important observation is that the nuclei configuration $\mathbf{R}$ changes only slightly from one time step to the next. Therefore a natural initial guess is to take the converged density matrix at the previous time step. This works better than a random initial guess but can still be improved. Various initial guess

strategies have been proposed in the literature. Examples include the core guess, which uses the density matrix obtained by diagonalizing the core Hamiltonian, the MinAO guess, where one solves the problem first on a minimal basis and projects onto a larger basis, and the superposition of atomic densities.

In all the approaches mentioned above, the initial guess only depends on the current configuration; no information from previous density matrices is used. However, in the context of molecular dynamics, many initial guess strategies do exploit such temporal information. The reference method is the extended Lagrangian approach [113], in which an auxiliary density is propagated in a time-reversible manner and used as the initial guess for the SCF procedure. This method is analyzed mathematically in [7]. This strategy is particularly effective, as it combines accurate guesses with excellent stability, often reducing the number of SCF iterations from several tens to just a few. Our aim in this work was to develop an alternative extrapolation method that further reduces the number of SCF iterations.

4.2.2 Time-reversible Grassmann extrapolation

When a property is expected to vary smoothly from one configuration to the next, it is tempting to predict its value at a new point as a linear combination of previously computed values. Unfortunately, the set of density matrices does not form a vector space but a manifold; a generic linear combination of two valid density matrices usually lies outside this set. To circumvent this, we work in the tangent space to the manifold at a reference point (see Figure 4.1), and provide linear combinations of initial guesses in this vector space.

The major issue is then how to choose the weights in that linear combination. We resolve this by using atomic descriptors, i.e. numerical fingerprints of a structure that encode its local chemical environment. By correlating the change in density matrix with these descriptors, we obtain a set of coefficients that produce an accurate extrapolation. The resulting scheme delivers highly reliable initial guesses, typically reducing the number of self-consistent field iterations compared with the extended Lagrangian approach. The method has been described in detail in our previous works [A26, A25, A24], and in the following we provide a few key points and results.

In this section, we adopt the notation related to (1.2.13), adding the nuclei configuration dependency $\mathbf{R}$. The number of electrons (or electron pairs) is N_{el}, and N is the dimension of the discretization space, $C_{\mathbf{R}} \in \mathbb{R}^{N\times N_{\mathrm{el}}}$ denotes the coefficients of the occupied orbitals, $D_{\mathbf{R}} \in \mathbb{R}^{N\times N}$ denotes the density matrix, and $\Lambda_{\mathbf{R}} \in \mathbb{R}^{N_{\mathrm{el}}\times N_{\mathrm{el}}}$ the diagonal matrix containing the energy levels. Further, $F_{\mathbf{R}}$ denotes the DFT-operator acting on the density matrix and $S_{\mathbf{R}}$ the overlap matrix of the basis functions. The modified coefficient matrix $\widetilde{C}_{\mathbf{R}} \coloneqq S_{\mathbf{R}}^{\frac{1}{2}} C_{\mathbf{R}}$ belongs to the Stiefel manifold defined as follows

$$\mathcal{S}t(N_{\mathrm{el}}, N) \coloneqq \left\{ V \in \mathbb{R}^{N\times N_{\mathrm{el}}} \;\middle|\; V^T V = \mathrm{Id}_{N_{\mathrm{el}}} \right\}, \tag{4.2.1}$$

due to the second equation in (1.2.13). In consequence the normalized density matrix $\widetilde{D}_{\mathbf{R}} = \widetilde{C}_{\mathbf{R}} \widetilde{C}_{\mathbf{R}}^T = S_{\mathbf{R}}^{\frac{1}{2}} D_{\mathbf{R}} S_{\mathbf{R}}^{\frac{1}{2}}$ belongs to the following set

$$\mathcal{M}_{\mathrm{Gr}} \coloneqq \left\{ D \in \mathbb{R}^{N\times N} \;\middle|\; D^2 = D,\ D^T = D,\ \mathrm{Tr}\, D = N_{\mathrm{el}} \right\}, \tag{4.2.2}$$

which can be identified with the Grassmann manifold of N_{el}-dimensional subspaces of $\mathbb{R}^N$ by means of the spectral projectors. For any $D \in \mathcal{M}_{\mathrm{Gr}}$, one can associate the tangent space $\mathcal{T}_D$ which has the structure of a vector space. The evolution of the electronic structure can therefore be seen as a trajectory $t \mapsto D_{\mathbf{R}(t)}$ on $\mathcal{M}_{\mathrm{Gr}}$ where $t \mapsto \mathbf{R}(t)$ denotes the trajectory of the nuclei. We then use that there exists a locally bijective mapping between the Grassmann

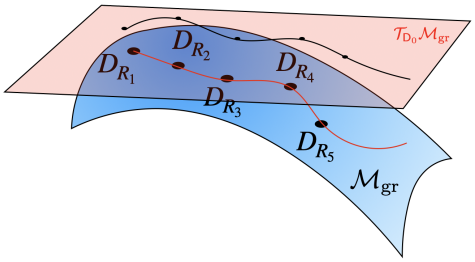


Figure 4.1: Dynamics on the Grassmann manifold and corresponding points on the tangent space.

manifold and its tangent space called exponential denoted by $\mathrm{Exp}_{\mathrm{Gr}}$ and its inverse is called the logarithm $\mathrm{Log}_{\mathrm{Gr}}$:

$$\mathrm{Log}_{\mathrm{Gr}} : \mathcal{M}_{\mathrm{Gr}} \to \mathcal{T}_{D_0}\mathcal{M}_{\mathrm{Gr}}, \qquad \mathrm{Exp}_{\mathrm{Gr}} : \mathcal{T}_{D_0}\mathcal{M}_{\mathrm{Gr}} \to \mathcal{M}_{\mathrm{Gr}},$$

where D_0 is a reference point. A key advantage of the Grassmann manifold is that these maps are explicit and computationally efficient, requiring only the thin singular value decomposition (SVD) of the coefficient matrix $C_{\mathbf{R}}$. Therefore the theory is expressed with the large density matrix $D_{\mathbf{R}}$, but in practice, the calculations are done on the thin coefficient matrix $C_{\mathbf{R}}$. We propose a linear approximation in the tangent space, as illustrated on Figure 4.1, that is

$$D(\mathbf{R}) \simeq \mathrm{Exp}_{\mathrm{Gr}} \left(\sum_{k=1}^{K} c_k(\mathbf{R}) \, \mathrm{Log}_{\mathrm{Gr}}(D_{\mathbf{R}_k}) \right).$$

The remaining challenge is to identify a suitable parameterization for the coefficients $c_k(\mathbf{R})$. To address this, we explored several strategies. In the first contribution [A26], we used Lagrange polynomials to construct the extrapolation. The approach led to encouraging results, but was restricted to situations where the explicit normal mode coordinates of the molecule are available; an assumption that does not hold in BOMD. For this, we turned to so-called descriptors, which are widely used in the context of interatomic potentials, as will be presented in Chapter 5. These descriptors are constructed so that they respect the relevant symmetries of the problem; for example, they are invariant under a rigid translation of the entire molecule. Indeed, in the context used for this work, that is molecules discretized with localized basis sets, the density matrix is also invariant with respect to this transformation. If the descriptors $d_{\mathbf{R}}$ are approximated as

$$d_{\mathbf{R}} \simeq \sum_{k=1}^{K} c_k d_{\mathbf{R}_k},$$

where $d_{\mathbf{R}_k}$'s are the descriptors for the configurations $\mathbf{R}_1$ to $\mathbf{R}_K$ then the density matrix on the tangent space can be well approximated as

$$\mathrm{Log}_{\mathrm{Gr}} D_{\mathbf{R}} \simeq \sum_{k=1}^{K} c_k \mathrm{Log}_{\mathrm{Gr}}(D_{\mathbf{R}_k}). \tag{4.2.3}$$

We therefore find the coefficients $(c_k(\mathbf{R}))_{k=1}^{K}$ by solving the regularized least square problem

$$\min_{\mathbf{c}\in\mathbb{R}^K} \left\| d_{\mathbf{R}} - \sum_{k=1}^{K} \mathbf{c}_k d_{\mathbf{R}_k} \right\|^2 + \alpha \|\mathbf{c}\|^2, \tag{4.2.4}$$

where α is a small real parameter. For the descriptors, we used the Coulomb matrix, which is one of the simplest possible descriptors defined by $C_{ij} = \frac{Z_i Z_j}{|\mathbf{R}_i - \mathbf{R}_j|}$ and provided very satisfactory results.

The idea behind the descriptor choice is that using two explicit nonlinear mappings, the remaining mapping to approximate becomes simpler. Hence the mapping from configuration space to density matrix is decomposed as the composition of three maps

$$\begin{aligned} \mathbb{R}^{3M} \to \mathcal{D} \ &\to \mathcal{T}_{D_0}\mathcal{M}_{Gr} \to \mathcal{M}_{Gr} \\ \mathbf{R} \mapsto d_{\mathbf{R}} &\mapsto \Gamma_{\mathbf{R}} \qquad\quad \mapsto D_{\mathbf{R}} = \mathrm{Exp}_{Gr}\left(\Gamma_{\mathbf{R}}\right), \end{aligned}$$

which are (1) Embedding of atomic positions in high-dimensional space of descriptors. This is an explicit mapping. (2) Map from descriptors to tangent space. This is where the approximation is done. (3) Grassman exponential. This is also an explicit mapping. To conclude, the final algorithm proposed in this method is simple and can be decomposed in the following steps. For a new configuration $\mathbf{R}$:

1. Compute the vector of descriptors $d_{\mathbf{R}}$,
2. Solve the least squares problem involving the descriptors (4.2.4),
3. Compute the linear combination on the tangent space (4.2.3),
4. Take the exponential to obtain the new density matrix.

System	N_{QM}	N_{MM}	N
OCP	129	4915	1038
APPA	31	16449	309
DMABN	21	6843	185
3HF	28	15018	290

Table 4.1: Overview of the system size in terms of number of QM-atoms (N_{QM}), number of MM-atoms (N_{MM}) and the total number of (QM) basis functions (N).

In [A25], this method has been tested for very large chemical systems, namely 3-hydroxyflavone (3HF) in acetonitrile, chromophores embedded in a biological matrix (OCP and APPA), as well as dimethylaminobenzonitrile (DMABN) in methanol. The main characteristics of the systems used for testing are given in Table 4.1. The numerical results were very satisfying, as our method denoted by G-Ext is always better than the extended Lagrangian method, with or without McWeeny (MW) purification. We report the results in Table 4.2.

A shortcoming of the initial implementation was that the total energy was not conserved when using the new initial guess (see the red plot in Figure 4.2), a property usually called time reversibility and one of the main features of the extended Lagrangian approach. We addressed this issue in a following article [A24]. The main idea was to symmetrize the initial guess with respect to time so that the dynamics becomes quasi time-reversible. This means that we impose the coefficients to be equal two by two. To be more precise, if n is the current time step of the molecular dynamics, and we are given q previous snapshots $\Gamma_{n-i} = \mathrm{Log}_{\mathrm{Gr}}(D_{n-i})$, for $i = 1, \ldots, q$, the approximation of the density matrix on the tangent space is written as

$$\widetilde{\Gamma}_n = -\Gamma_{n-q} + \sum_{i=1}^{\widetilde{q}} c_i \left(\Gamma_{n-i} + \Gamma_{n-q+i}\right), \tag{4.2.5}$$

Table 4.2: Performances of the G-Ext method for different number of extrapolation points, compared with the XLBO algorithm with and without McWeeny purification. All the results were obtained using a 10^{-5} convergence threshold for the root-mean-square increment of the density matrix and are derived from a picosecond long molecular dynamics simulation, using a 0.5 femtosecond time step. We report on the average number of iterations required to converge the SCF, together with the associated standard deviation. Note that the first 8 steps were discarded.

	OCP		DMABN		APPA		3HF	
Method	Average	σ	Average	σ	Average	σ	Average	σ
XLBO	3.82	0.66	3.98	0.16	3.00	0.03	4.00	0.14
XLBO-MW	2.95	0.31	3.76	0.56	3.00	0.34	3.96	0.31
G-Ext(3)	2.57	0.84	3.54	0.78	2.95	0.50	3.09	0.41
G-Ext(4)	2.48	0.88	3.14	0.62	2.51	0.50	3.25	0.68
G-Ext(5)	2.25	0.96	3.23	0.75	2.51	0.50	3.30	0.72
G-Ext(6)	2.20	0.96	2.99	0.02	2.51	0.50	3.14	0.56

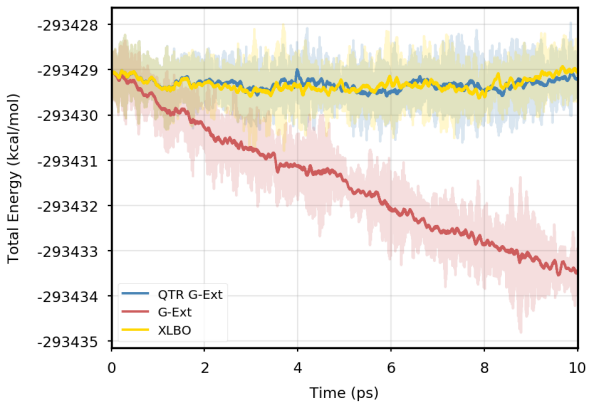


Figure 4.2: Total energy as a function of simulation time for DMABN, using a 10^{-5} convergence threshold for the SCF.

where $\widetilde{q} = q/2$ if q is even, while $\widetilde{q} = (q-1)/2$ if q is odd. The coefficients c_i are computed in a similar way as before, i.e. by solving the least-squares problem with Tikhonov regularization

$$\min_{\mathbf{c}\in\mathbb{R}^{\widetilde{q}}} \left\{ \left\| d_{\mathbf{R}_n} + d_{\mathbf{R}_{n-q}} - \sum_{i=1}^{\widetilde{q}} c_i \left(d_{\mathbf{R}_{n-i}} + d_{\mathbf{R}_{n-q+i}} \right) \right\|^2 + \alpha \left\| \mathbf{c} \right\|^2 \right\}.$$

In terms of the number of iterations, the results were even better than the ones obtained with the non time-reversible method, and the long-term energy conservation is increased, as shown in Figure 4.2 (blue plot).

In conclusion, we have managed to reduce significantly the computation cost of BOMD by providing accurate initial guesses for the density matrices based on a combination of a tangent space structure and the use of molecular descriptors, that allow, moreover, for a time-reversible structure.

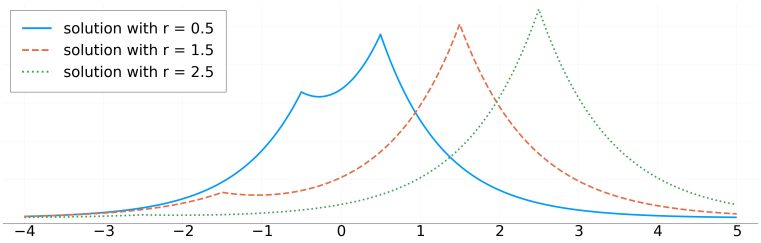


Figure 4.3: Three example solutions in $\mathcal{M}$.

4.3 Nonlinear reduced order model based on optimal transport

We now present the main aspects of [S1], which presents a nonlinear reduced order model for a toy problem based on mixture Wasserstein barycenters. The first contribution is to estimate the decay rate of the Kolmogorov n-width of the set of solutions in several settings, including the standard L^2-norm as well as Wasserstein and mixture Wasserstein distances. It is shown that the decay rate is higher for optimal transport-based distances than for the L^2-norm. The second contribution is a nonlinear reduced basis method based on an offline greedy algorithm, and an efficient stochastic energy minimization in the online phase, for which numerical results are very encouraging towards 3D applications.

4.3.1 A toy problem in electronic structure

Ultimately, we are interested in the electronic structure problems presented in Chapter 1 but for the sake of simplicity, in [S1], we studied a simpler, one-dimensional toy problem. We consider the eigenvalue partial differential equation parameterized by the nuclei positions $\mathbf{R} := (r_1, \dots, r_M) \in \mathbb{R}^M$ with charges $\mathbf{z} := (z_1, \dots, z_M) \in (\mathbb{R}_+^*)^M$ for $M \in \mathbb{N}^*$. More precisely, we are looking for the lowest eigenvalue $E_{\mathbf{R}} \in \mathbb{R}$ and corresponding eigenstate $\Phi_{\mathbf{R}} \in H^1(\mathbb{R})$ satisfying

$$-\frac{1}{2}\Phi_{\mathbf{R}}'' + \left(-\sum_{m=1}^{M} z_m \delta_{r_m}\right)\Phi_{\mathbf{R}} = E_{\mathbf{R}}\Phi_{\mathbf{R}}. \tag{4.3.1}$$

This problem can be seen as the electronic structure problem of finding the ground state of an Hamiltonian of the form $-\frac{1}{2}\Delta + V$, with a potential V taken as a sum of Dirac masses $V := -\sum_{m=1}^{M} z_m \delta_{r_m}(x)$ localized at atomic positions $\mathbf{R}$ with charges $\mathbf{z}$. The choice of this simple Hamiltonian relies on the nice property that there exists a unique strictly positive eigenvector solution to this problem, which we normalize in L^1-norm in order to associate it with a probability distribution and it is explicitly given by [119, Section 3.1]

$$\Phi_{\mathbf{R}} = \sum_{m=1}^{M} \pi_m^{\mathbf{R},\mathbf{z}} \mathrm{S}_{\zeta_{\mathbf{R},\mathbf{z}}, r_m}, \tag{4.3.2}$$

with $\boldsymbol{\pi}^{\mathbf{R},\mathbf{z}} = \left(\pi_m^{\mathbf{R},\mathbf{z}}\right)_{m=1}^{M} \in (\mathbf{R}_+)^M$ of total sum equal to 1, $\zeta_{\mathbf{R},\mathbf{z}} > 0$, and where for all $\zeta > 0$ and all $r \in \mathbb{R}$, the Slater function $\mathrm{S}_{\zeta,r}$ is defined by

$$\mathrm{S}_{\zeta,r} : x \longmapsto \frac{\zeta}{2} e^{-\zeta|x-r|}. \tag{4.3.3}$$

In Figure 4.3, we plot three solutions with $M = 2$ with different interatomic spacing.

4.3.2 Offline greedy algorithm

We first present the greedy algorithm used in the offline phase. Let $\mathbf{z} \in \left(\mathbb{R}_+^*\right)^M$ be fixed positive charges, $\mathcal{M}_{\mathbf{z}}^I = \{\Phi_{\mathbf{R}},\ \mathbf{R} \in I^M\}$ with $I \subset \mathbb{R}$ an interval be a set of solutions of (4.3.1) and $\mathcal{M}_{\text{train}} \subset \mathcal{M}_{\mathbf{z}}^I$ a finite training set of already computed solutions called snapshots. The aim here is to select the most representative snapshots in $\mathcal{M}_{\text{train}}$, so that any solution $u \in \mathcal{M}_{\mathbf{z}}^I$ can be efficiently approximated with only a few snapshots. At each iteration, we select the snapshot in $\mathcal{M}_{\text{train}}$ for which the approximation as a mixture barycenter (3.2.3) denoted by $\text{Bar}_{\text{MW}_2}^{\boldsymbol{t}}(m^1, \ldots, m^n)$ of previously selected snapshots leads to the largest error, as presented in Algorithm 1.

For simplicity we decompose the solutions $\Phi_{\mathbf{R}} \in \mathcal{M}_{\text{train}}$ as mixtures $m = \sum_{k=1}^{K} \pi_k m_k$. These mixtures being solutions to problem (4.3.1) means that their elements, denoted by m_k are Slater functions with a scale parameter ζ independent of k and position parameter r_k. The space Ω_n is larger than the usual Λ_n but still guarantees that the barycenters are well-defined.

Algorithm 1 Greedy algorithm

Input: $\mathcal{M}_{\text{train}}$, training set; N, number of elements to select
Select m^1 and m^2 solutions to $\underset{(m^1,m^2)\in\mathcal{M}_{\text{train}}}{\text{argmax}} \text{MW}_2(m^1, m^2)^2$.
$\mathcal{B} := \{m^1, m^2\}$
for $n = 2, \ldots, N-1$ **do**
Select

$$m^{n+1} \in \underset{m\in\mathcal{M}_{\text{train}}}{\text{argmax}} \min_{\boldsymbol{t}\in\Omega_n} \text{MW}_2\left(m, \text{Bar}_{\text{MW}_2}^{\boldsymbol{t}}(m^1, \ldots, m^n)\right)^2, \tag{4.3.4}$$

where

$$\Omega_n = \left\{ \boldsymbol{t} \in \mathbb{R}^n,\ \sum_{i=1}^{n} \frac{t_i}{\zeta^i} > 0 \right\}.$$

$\mathcal{B} = \mathcal{B} \cup \{m^{n+1}\}$
end for
Output: Reduced basis $\mathcal{B} \subset \mathcal{M}_{\text{train}}$

4.3.3 Online optimization algorithm

Once the best snapshots are selected with the greedy algorithm, we want to efficiently compute approximations of solutions, given a new position for the nuclei $\mathbf{R}$. For this we consider the minimization problem associated with the eigenvalue problem (4.3.1), that is we minimize the energy of the new system over the set of barycenters of selected snapshots. More precisely, assume that we selected N mixture solutions $m^1, \ldots, m^N$ in the offline phase and we want to obtain an approximation to the solution with molecular parameters $\mathbf{R}$. We consider the following optimization problem

$$\min_{\boldsymbol{t}\in\Omega_N} E_{\mathbf{R}}\left(\text{Bar}_{\text{MW}_2}^{\boldsymbol{t}}(m^1, \ldots, m^N)\right), \tag{4.3.5}$$

where Ω_N is the extended set of admissible barycenters, as in the greedy algorithm, and $E_{\mathbf{R}}$ is the energy associated with the eigenvalue problem (4.3.1). Note that the energy can easily be computed, but is nonconvex as a function of $\boldsymbol{t}$, and exhibits many local minima. Therefore solving problem (4.3.5) requires to use a global optimization algorithm, preferably very robust

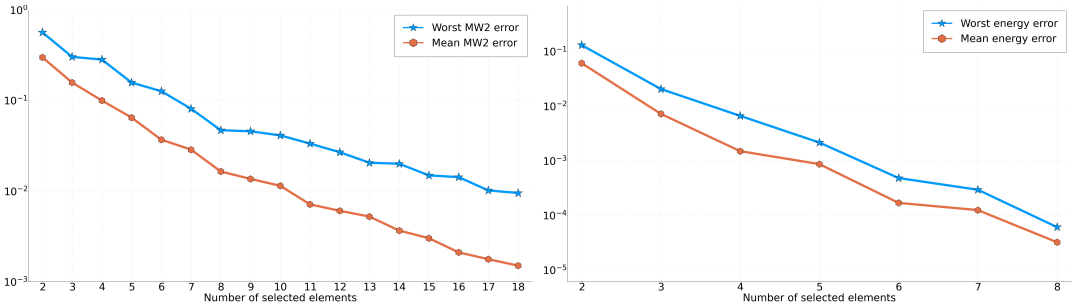


Figure 4.4: Left: decay of the projection error in the offline phase; right: decay of the energy error in the online phase, for 51 equally distributed elements for $r \in [0.5, 3]$.

to ensure that the global minimizer is found. As described in Algorithm 2, we use a quasi-Newton minimization algorithm (LBFGS) with evenly distributed starting points using a Sobol sequence on a representative set $B_N = [-B, B]^N \cap \Omega_N$ of values of $\boldsymbol{t}$.

Algorithm 2 ONLINE OPTIMIZATION

Input: Reduced basis $m^1, \ldots, m^N$, $B_N = [-B, B]^N \cap \Omega_N$, starting points $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_L$ in B_N
for $l = 1, \ldots, L$ **do**
  Compute $\boldsymbol{t}_l^*, E_l^*$ minimizer and energy solution found
  by optimizing $E_{\mathbf{R}}\left(\mathrm{Bar}^{\boldsymbol{t}}_{\mathrm{MW}_2}(m^1, \ldots, m^N)\right)$ for $\boldsymbol{t} \in \Omega_N$
  with starting point $\boldsymbol{t}_l$ with a LBFGS algorithm
end for
Output: Minimizer $\boldsymbol{t}^* = \mathrm{argmin}_{l=1,\ldots,L} E_l^*$

4.3.4 Numerical results

The numerical results were performed with a system with two nuclei, i.e. $M = 2$, and with charges $\mathbf{z} = (0.8, 1.1)$ and $\mathbf{R} = (-r, r)$ for $r \in \mathbb{R}_+$. For the training set, the r's are equally distributed on the interval $[0.5, 3]$ with training set of size 251.

Both offline and online numerical results are very satisfactory. In Figure 4.4 (left), we plot the mean error on the training set (red) and the maximum error (blue), which decreases very fast, possibly exponentially. Moreover, we gain about two orders of magnitude between 2 and 15 added snapshots in the reduced basis on the mean error. In Figure 4.4 (right), we provide both the maximum error (blue) and the mean error (red) over a test set of 51 equally distributed elements for $r \in [0.5, 3]$. We observe that the energy maximum error decreases by three orders of magnitude from 2 to 8 snapshots, which is particularly encouraging. Moreover, the method allows one to extrapolate outside of the range of training data, as the formula for the barycenter stays valid in this context.

To conclude this section, this contribution is extremely promising toward nonlinear reduced models for electronic structure calculations based on mixture Wasserstein barycenters, as the mixture Wasserstein barycenters are valid in higher dimensions. A current drawback is the computational cost of solving the multi-marginal problem that arises in computations of barycenters, but several techniques [64, 140] may lead to a sharp decrease of the computational cost.

4.4 Approximating the density to pair-density map

In this section we are interested in the approximation of the map from the electronic density $\rho_{\mathbf{R}}$ to the pair density $\rho_{2,\mathbf{R}}$. The underlying problem of interest is the full Schrödinger equation (1.1.1). One reason why this map is of interest is that the electron-electron repulsion energy contribution

$$V_{ee}[\Psi_{\mathbf{R}}] = \int_{\mathbb{R}^{3N_{\mathrm{el}}}} \sum_{1\leq i<j\leq N_{\mathrm{el}}} v_{ee}(\mathbf{r}_i - \mathbf{r}_j)|\Psi_{\mathbf{R}}(\mathbf{r}_1, \dots, \mathbf{r}_{N_{\mathrm{el}}})|^2 d\mathbf{r}_1 \cdots d\mathbf{r}_{N_{\mathrm{el}}},$$

where v_{ee} is in general the Coulomb potential can be simply written with the pair density as

$$V_{ee}[\Psi_{\mathbf{R}}] = \int_{\mathbb{R}^{2d}} v_{ee}(\mathbf{r} - \mathbf{r}')\rho^{\Psi}_{2,\mathbf{R}}(\mathbf{r}, \mathbf{r}')d\mathbf{r}\, d\mathbf{r}'.$$

Accurate approximations of the pair density would therefore yield precise estimations of this interaction energy. In Kohn–Sham density functional theory, this term is approximated via exchange-correlation functionals, directly or indirectly as a functional of the electronic density. Developing reliable approximations of the interaction energy that remain accurate when standard functionals fail, such as when subsystems dissociate would be valuable. Additionally the pair density itself provides precious insight into correlation between pairs of electrons within the system.

A natural extension to [S1] would be to learn the density to pair-density map using mixture Wasserstein barycenters (3.2.3). This is currently work in progress. So far we have started to learn the density to pair-density map with a different approach relying on the theory of copulas in statistics, presented in [A18]. The main idea behind the copulas is to separate the marginals, which are in this context the electronic density $\rho_{\mathbf{R}}$ from the multidimensional dependence structure, i.e. the pair density $\rho_{2,\mathbf{R}}$. This is based on Sklar's theorem which we use here in the case of particles in 1D. The copula density $c_{\mathbf{R}}$ is defined on $[0,1]^2$ (up to a normalization constant) as

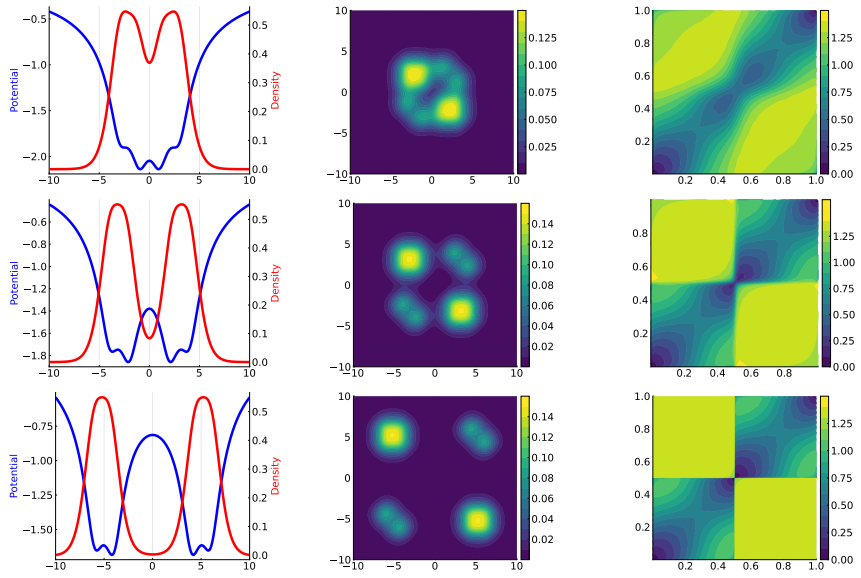
$$c_{\mathbf{R}}(r_1, r_2) = \frac{\rho_{2,\mathbf{R}}(F(r_1), F(r_2))}{\rho_{\mathbf{R}}(F(r_1))\rho_{\mathbf{R}}(F(r_2))}, \tag{4.4.1}$$

with $F(r) = \mathrm{icdf}(\rho_{\mathbf{R}})(r)$. It can be seen as an adequate change of variables which simplifies the approximation of $\rho_{2,\mathbf{R}}$ when the nuclei configuration change, especially at dissociation since the copula converges to a well-defined limit (see Figure 4.5).

In fact, the general theory of copulas does not apply directly to electronic structure problems in 3d, since in general it decomposes a multivariate distribution on $\mathbb{R}^{N_{\mathrm{el}}}$ in terms of its N_{el} univariate marginals on $\mathbb{R}$ and a copula. Instead, for electrons in $\mathbb{R}^3$, one should factorize the configuration space $\mathbb{R}^{3N_{\mathrm{el}}}$ into N_{el} factors $\mathbb{R}^3$ (instead of $3N_{\mathrm{el}}$ factors $\mathbb{R}$), and separate the N_{el}-electron distribution on $\mathbb{R}^{3N_{\mathrm{el}}}$, or the pair density on $\mathbb{R}^6$, in terms of a copula and its marginal on $\mathbb{R}^3$, which are given by the single-particle density $\rho_{\mathbf{R}}$.

In [A18], we therefore start by generalizing the theory of copulas to precisely such factorizations, using optimal transport theory [144, 63]. We then observe the copula structure for one-dimensional systems from 2 to 4 electrons, and we provide a simple fit for the copula which is asymptotically exact and provides excellent results for the approximation of the pair density.

More precisely we show theoretically and we observe numerically that the copula asymptotically looks like a checkerboard, with insets that look like copulas for systems with a smaller number of electrons. For example, in Figure 4.5 we observe the checkerboard structure for the dissociation of a system with four particles as two two-particle systems. We see that the copula becomes constant on the two off-diagonal parts, and that the diagonal parts correspond to the



(a) External potential V_{ext} and electronic density (b) Pair density (c) Copula (4.4.1)

Figure 4.5: Dissociation of four-particle system into two two-electron charged densities. Parameter a for top row: $a = 1$, second row: $a = 2$, bottom row: $a = 3$, with nuclei positions at $(-a-1.5, -a+1.5, a-1.5, a+1.5)$.

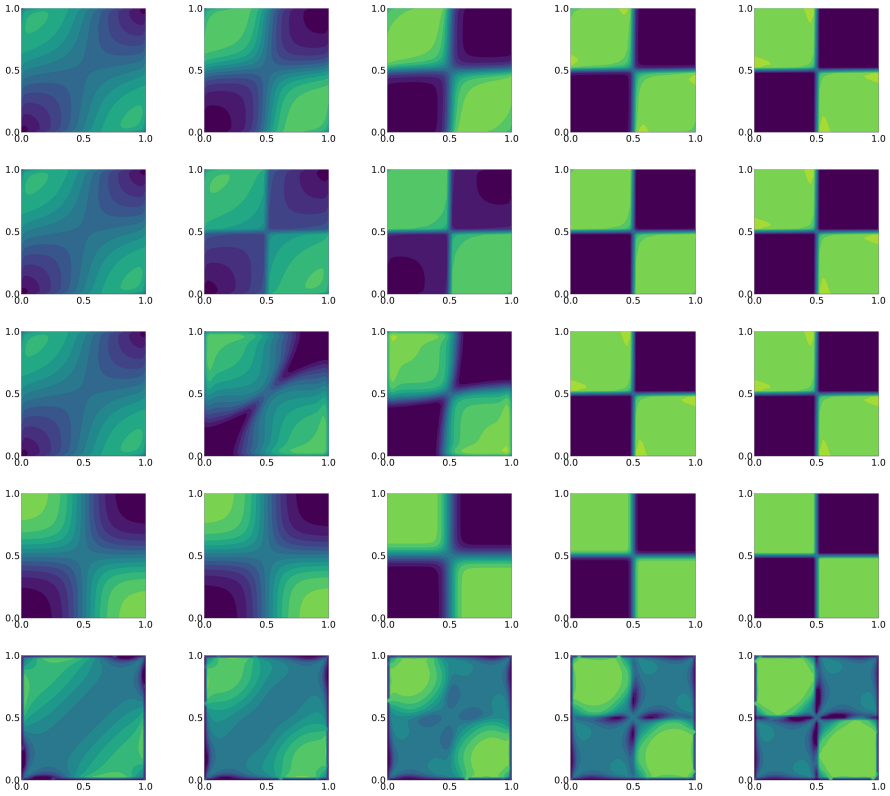
copula of the two-particle system of Figure 4.6. Also the values of the copulas on the constant parts match those predicted by theory.

Once we have understood the asymptotic structure of the copula at dissociation, we propose a few data-driven generalized copula models, trained on suitable reference pair densities with asymptotic results built in as exact constraints. We hope this to be a way to propose new DFT models which remain accurate in strongly correlated regimes. The results of the proposed fitting procedures between two molecular configurations are displayed on Figure 4.6. We approximate the intermediate copulas for two particles, i.e. for $a = 1.5, 2, 2.5$, only from the knowledge of the copulas for $a = 1$ and $a = 3$. The different proposed models for the copula are: linear, Wasserstein barycenter, small neural network with a self-attention layer. The most promising result is given by the one-layer neural network. In any case, all fitted models perform way better than the local density approximation [94], which is a standard model in DFT. To conclude the generalized copula theory is very promising for further three-dimensional tests with electronic structure calculations.

4.5 Perspectives

I now present research perspectives related to nonlinear reduced order models.

- **Grassmann extrapolation for materials systems** A natural extension of the Grassmann extrapolation technique would be to look at periodic systems in order to reduce the cost of molecular dynamics simulations and geometry optimization. This requires to tackle two main difficulties. First, the extrapolation technique was particularly successful because localized bases are used for molecules, which implies that the coefficients of the density matrices in such basis do not change too much when the molecular configuration $\mathbf{R}$ changes.



(a) $a = 1$ (b) $a = 1.5$ (c) $a = 2$ (d) $a = 2.5$ (e) $a = 3$

Figure 4.6: Copulas for two-particle systems obtained from different models. Exact (top row), Linear interpolation (second row), Wasserstein barycenter (third row), one-parameter neural net (fourth row), LDA (bottom row). Pictures furthter to the right correspond to a larger internuclear distances. The color limits are (0,2.5).

For periodic systems, the discretization bases are often plane waves so that nothing guarantees that the coefficients slowly vary. This is the main change between molecules, for which the method was successfully used, and periodic insulators. For this, we should first test whether the extrapolation works with delocalized basis functions. If, as we expect, this is not the case, we could look at ways to generate localized bases from nonlocal bases functions, such as Wannierization [105]. We expect this to work reasonably well for insulators, considering calculations at a single $\mathbf{k}$-point. This may require to modify the Wannierization procedure in order to obtain basis functions which smoothly vary with the configuration variable $\mathbf{R}$, and not only the $\mathbf{k}$-points. One solution for this might be to attach the basis functions to specific atoms.

To go beyond insulators and consider metallic systems, a crucial difference appears. Indeed, the density matrix is no longer an orthogonal projector, therefore the density matrix does not belong to the Grassmann manifold. More precisely, the density matrix writes

$$D(\mathbf{R}) = \sum_{i=1}^{\tilde{N}} \alpha_i(\mathbf{R}) \phi_{\mathbf{R},i} (\phi_{\mathbf{R},i})^T,$$

for some $\alpha_i(\mathbf{R}) \geq 0$ called occupation numbers and orthonormal vectors $\phi_{\mathbf{R},i}$. Therefore, while the density matrix itself does not belong to the Grassmann manifold, the functions

$\phi_{\mathbf{R},i}$ belong to a Stiefel manifold. Exploring the tangent space of Stiefel manifolds [2], one could derive a similar extrapolation formula, which might, however, be more costly to compute due to the possibly high number of functions $\phi_{\mathbf{R},i}$ in the expansion. Moreover, in that setting, we also need a good strategy for extrapolating the coefficients $\alpha_i(\mathbf{R})$. We could start with simple interpolation techniques (e.g. linear expansion) before moving to nonlinear representations if necessary. A further step would be to directly consider the manifold of metallic density matrices, and derive exponentials and logarithms. We could first test the method on systems with varying lattice size, i.e. a case where the varying parameter $\mathbf{R}$ is clearly identified, as a toy model.

- **Theoretical explanations of the Grassmann extrapolations** So far, explanations why the Grassmann extrapolation works so well are lacking. A first hint is given by the multipoint formula presented in Chapter 2. However this is not sufficient. Therefore, it would be interesting to understand how the choice of the descriptors plays a role in the approximation. Also, it would be key to mathematically understand the time-reversibility property and write down carefully the conditions that lead to a stable dynamics. This would hopefully lead to understanding why the quasi time-reversible method provides such amazing results, and possibly improving them further.

- **Theoretical results for nonlinear reduced order modeling** At that point, apart from some estimates for nonlinear Kolmogorov n-widths for simple equations, no general theoretical convergence rates are available in this nonlinear "metric" context. For instance, it would be of interest to extend existing theoretical results that relate the convergence of the Kolmogorov n-width to the convergence behavior of the greedy algorithm, similarly to what is done in [27] for projection-based model order reduction.

- **Error bounds for nonlinear reduced order modeling** We could also develop error bounds for the reduced order methods based on optimal transport, first focusing on the classical optimal transport extrapolations such as those provided in [S1]. We could first provide an *a priori* estimation, i.e. provide a rate of convergence of the error with respect to the number of selected solutions, and then move to *a posteriori* error estimation. This would be a first step in deriving *a priori* and *a posteriori* estimates in metric spaces, where such bounds are, to the best of my knowledge, not available. In the case of the toy model considered in [S1], the fact that the problem is one dimensional makes it possible to recast the metric space estimation as a Hilbert space estimation as in [55], and use the inverse cumulative distribution function of the solution as a change of variable. We then hope to be able to treat more general models, in particular first linear eigenvalue problems in 3d, and then nonlinear problems. If this works out well, we could perform a similar analysis on Quantum Optimal Transport extrapolations.

- **Density to pair-density map** A natural extension of what has been presented in this chapter would be to provide a good approximation of the density to pair-density map for 3D calculations. Indeed, with the mixture Wasserstein distance, the marginal-constrained barycenters presented in Chapter 3, and the greedy algorithm presented in this chapter, we are ready to perform such calculations.

Chapter 5

Data-driven methods: interatomic potentials to Hamiltonian models

This chapter deals with data-driven methods for approximating quantities of interest, typically energy and forces, as functions of the atomic configuration $\mathbf{R}$ of the system. The main difference compared to the previous chapters is that the considered quantities of interest are directly fitted using a database. This type of machine-learning approach has exploded over the past 10-20 years, and is now extensively used in interatomic potential modeling.

There are three main ingredients in these methods: the data, the parameterization, and the cost function used in the training process. To start this chapter, we describe the problem of learning the potential energy surface (PES) in Section 5.1 through these three aspects. The parameterization used for learning the PES often relies on molecular descriptors, which capture atomic environments as vectors. Moreover the PES needs to respect some symmetry, which is ensured through symmetry of the descriptors. A review of available descriptors is presented in Section 5.2, with a particular emphasis on Permutation-Invariant Potentials (PIPs) and the Atomic Cluster Expansion (ACE) which I contributed to [A27, A1, A15, A2]. To open the discussion, we present in Section 5.3 recent advances in using data-driven approaches for learning electronic structure-related quantities directly. We finally present some perspectives related to this subject in Section 5.4.

5.1 Data-driven approach

5.1.1 Quantities of interest and data

In this chapter, the main motivation is to directly learn a map $\mathbf{R} \mapsto E(\mathbf{R})$ where E is some quantity of interest, i.e any property related to quantum chemistry. The most popular choice is the potential energy and the interatomic forces. Other examples are derived quantities, such as the band spectrum, the dipole moment, or more involved quantities such as the wavefunction, the Hamiltonian, the electronic density, and density matrix.

Data-driven approaches require a dataset generated using a reference method, ideally one of high accuracy. In the context of interatomic potentials, this is a major advantage: if the database is constructed from electronic structure calculations, typically DFT, but also potentially more accurate wavefunction-based methods, then, provided the learned energy functional generalizes well beyond the training data, one can achieve energy and force predictions with near-quantum accuracy at a fraction of the computational cost.

5.1.2 Parametrization

Compared to reduced order modeling approaches, machine-learning methods do not allow for online calculations such as solving linear systems. Instead, the mapping $\mathbf{R} \mapsto E(\mathbf{R})$, is approximated by a surrogate model $\mathbf{R} \mapsto E_\theta(\mathbf{R})$, constructed entirely during the offline phase, referred to as training in the machine-learning (ML) community. Then, during the online phase, i.e., when the quantity of interest must be evaluated for a new atomic configuration $\mathbf{R}$, the approximation $E_\theta(\mathbf{R})$ is simply evaluated without further adaptation or computation. The general approach is to select a functional form for $E_\theta(\mathbf{R})$ that depends on a set of parameters θ, which can be adjusted so that the function fits the data at known points. Different ML models correspond to different choices of the functional form E_θ.

In learning the PES, a key aspect is the decomposition of the map E_θ into two parts as

$$E_\theta : \begin{cases} \mathcal{C} \to \mathbb{R}^{N_d} \to \mathbb{R} \\ \mathbf{R} \mapsto d_{\mathbf{R}} \mapsto E_\theta(d_{\mathbf{R}}), \end{cases} \tag{5.1.1}$$

where $\mathcal{C}$ is the set of atomic configurations, N_d is the size of the descriptor space. The first map $\mathbf{R} \mapsto d_{\mathbf{R}}$ is called the descriptor map and transform the Cartesian coordinates of the atoms into a suitable representation of size N_d. This allows to map configurations with arbitrary number of atoms to the same vector space, and hence account for systems of various sizes. Moreover the descriptor map takes care of the symmetries. The second part maps the descriptors to the output value. Since the input space is $\mathbb{R}^{N_d}$, standard functions can be used here, such as linear maps or neural networks. We present a short overview of available methods in the literature for the PES.

Linear methods

In linear methods the basis consists of the chosen descriptors. Since the descriptors often couple the information of a few particles at a time, the potential energy is written as a body-order expansion. This consists in decomposing first the energy as a sum of local components depending on atomic environments

$$E_\theta(\mathbf{R}) = \sum_i \widetilde{E}_\theta(\mathbf{R}_i; \mathbf{R} - \mathbf{R}_i). \tag{5.1.2}$$

Second the function $\widetilde{E}_\theta$ is written as a sum of terms coupling only k particles

$$\widetilde{E}_\theta(\mathbf{R}_i; \mathbf{R}-\mathbf{R}_i) = E_{\theta,1}(\mathbf{R}_i) + \sum_j E_{\theta,2}(\mathbf{R}_i; \mathbf{R}_j - \mathbf{R}_i) + \sum_{j,k} E_{\theta,3}(\mathbf{R}_i; \mathbf{R}_j - \mathbf{R}_i, \mathbf{R}_k - \mathbf{R}_i) + \dots \tag{5.1.3}$$

The hope is that this series converges fast with respect to the maximal body order. Then one way to ensure that the potential energy is symmetric with respect to rotation and permutation of identical atoms is that the functions $E_{\theta,1}, E_{\theta,2}, \dots, E_{\theta,K}$ satisfy these symmetries too. This is where descriptors come into play. The body-order expansion also helps the learning process as $E_{\theta,i}$ is considered local: terms involving atoms that are far away do not contribute. This also helps transferability, i.e. increases the accuracy outside the training set and for different molecular systems, and ensures the extensivity of the predictions.

Moreover, this expansion transforms the problem of approximating a function with a potentially infinite number of particles such as in the case of materials, to functions with a finite number and even reasonably small number of variables. The main advantage of the linearity is that it underlies the vast majority of empirical force fields, and training is usually cheap

compared to nonlinear methods. It often also offers a better explainability than nonlinear models. As for the evaluation cost, it is directly linked to the computational cost of evaluating the descriptors.

Linear methods include the Moment Tensor Potential (MTP) [136], the Atomic Cluster Expansion (ACE) [52], or the Spectral Neighbor Analysis Potential (SNAP) [142].

Nonlinear methods

Nonlinear methods are also used for fitting interatomic potentials. One of the first neural networks for this application was proposed by Behler and Parinello [21]. In this work, the neural network was simple and consisted of two fully connected hidden layers with 40 nodes each, that is about thousands parameters in total. One advantage is the very low evaluation cost, and the fact that the evaluation cost does not depend on the amount of training data points. Other methods use feed forward neural networks, such as the ANI [137] and DeepMD [146], but these neural networks are way larger than the Behler–Parinello network.

Another type of nonlinear architecture consists of kernel methods, mostly used in the Gaussian Approximation Potential (GAP) [12]. The main idea is to define a similarity measure between atomic environments as a L^2 scalar product between densities integrated over rotations. Numerically speaking, solving the optimization problem only requires to solve a linear system. As a drawback the evaluation cost depends on the number of configurations in the database.

More recently message-passing neural networks have been highly developed for interatomic potentials. They are based on different types of descriptors, possibly invariant or equivariant. Among the most popular are MACE [15], Schnet [134], and NequIP [16].

5.1.3 Cost functions and training

Once a functional form G_θ has been chosen, one needs to fit the parameters θ to match the data. This is done by minimizing a cost function. A typical cost function is the mean square error on the data: if $(\mathbf{R}^{(j)})_{i=1}^{N_{\text{train}}}$ are the data points for which we have already computed the exact function $E(\mathbf{R}^{(j)})$, the cost to minimize is

$$J(\theta) = \sum_{j=1}^{N_{\text{train}}} \|E(\mathbf{R}^{(j)}) - E_\theta(\mathbf{R}^{(j)})\|^2,$$

where the norm is adapted to the quantity to compute, typically absolute value for the energy and Euclidean norm for the forces.

When one has several types of data in the database, typically energy and forces, then one can combine the two types of data. E.g. if $E(\mathbf{R})$ denotes the potential energy, and $F(\mathbf{R})$ the atomic forces, then one would typically take

$$J(\theta) = \sum_{j=1}^{N_{\text{train}}} \alpha_E \|E(\mathbf{R}^{(j)}) - E_\theta(\mathbf{R}^{(j)})\|^2 + \alpha_F \|F(\mathbf{R}^{(j)}) - F_\theta(\mathbf{R}^{(j)})\|^2,$$

where α_E and α_F are tunable positive parameters that indicate the relative importance of the two quantities.

The minimization of the cost function, also called training, is done with an optimization method, which depends on the model G_θ. If G_θ is linear in its parameters, then the problem is a linear least square problem, which, depending on the size of the problem, can be solved either directly or with iterative methods. If G_θ is nonlinear with respect to its parameters, then typically stochastic gradient methods are used, such as the Adam algorithm [92] or quasi-Newton methods, depending on the number of parameters.

A regularization term is often added to the cost function such as the L^2-norm of the parameters θ,

$$J(\theta) = \sum_{i=1}^{N_{\text{train}}} \|E(\mathbf{R}^{(j)}) - E_\theta(\mathbf{R}^{(j)})\|^2 + \varepsilon\|\theta\|_2^2,$$

where ε is a small positive parameter to choose. This is called Tikhonov regularization, and penalizes large values of the parameters. This often improves the extrapolation capabilities outside of the database, avoiding overfitting.

Another common regularization technique is to add the L^1-norm of the parameters, also known as Lasso regularization. On the one hand, for linear parametrization, it increases the complexity of training, as the optimization problem becomes nonlinear. On the other hand, it promotes sparsity in the solution, meaning that only a few parameters are expected to be nonzero. This sparsity can, depending on the chosen functional form, be exploited to obtain a fitted model that is faster to evaluate. Many other penalizations can be used to enforce known behavior of the function to fit. E.g. in [A27], we penalized the Laplacian of the potential energy components in order to favor smooth components of the interatomic potential.

To summarize, the fitting procedure is used to construct an approximate function $E_\theta(\mathbf{R})$ using the data of the training set. The parameterization is then evaluated on a dedicated test set of configurations to assess how well it matches the true function. It can then be validated in more realistic settings, such as molecular dynamics simulations for interatomic potentials.

5.2 Atomic descriptors

In quantum chemistry and related fields, the functions to fit satisfy some physical constraints. Namely, they transform in a prescribed way when the nuclei configuration is changed according to some rigid-body motion, or permutation of the positions of identical atoms. This can be accounted for in the fitting procedure. One way to obtain such symmetry is to add data in the database for which we know the result due to the symmetry argument. But this is in general intractable since it highly increases the size of the database, even exponentially in the case of permutation invariance. The other choice, which is usually preferred, is to incorporate the symmetry directly in the functional E_θ, ensuring that the fitted function satisfies the required symmetry.

5.2.1 Symmetries

Permutation invariance

In electronic structure, as well as for interatomic potentials, identical particles are indistinguishable. This means that the values of the related quantities of interest are invariant with respect to a permutation of the position of identical atoms (in the classical context) or nuclei (in the quantum context). For the potential energy, this means that

$$E(\mathbf{R}) = E(\mathbf{R}_\sigma),$$

where σ denotes a permutation of identical atoms.

Rotation invariance and equivariance

On top of permutation invariance, there is also prescribed transformation with respect to rigid-body motion, that is inversion, translation, and rotations. In the case of the potential

energy, the function is invariant with respect to these transformations. In particular for any configuration $\mathbf{R}$, as well as any 3×3 matrix $Q \in O(3)$, there holds

$$E(Q\mathbf{R}) = E(\mathbf{R}),$$

where $Q\mathbf{R} = (Q\mathbf{R}_1, \dots, Q\mathbf{R}_M)$ rotates similarly all coordinates of the nuclei. For the forces though, the transformation is different as when the nuclei are rotated, the forces are rotated in the same way, that is

$$F(Q\mathbf{R}) = QF(\mathbf{R}).$$

More generally, such a transformation is called equivariance. Given a symmetry group $\mathcal{S}$, the equivariance property reads for a function G:

$$\forall g \in \mathcal{S},\ \forall \mathbf{R} \in \mathcal{C}, \quad G(g \cdot \mathbf{R}) = g \cdot G(\mathbf{R}),$$

where the first $\cdot$ is a group action on the configuration space, and the second $\cdot$ is a group action on the output space of G.

5.2.2 Construction of the descriptors

The atomic descriptors are designed to satisfy these symmetries. The translation invariance of the energy is automatically satisfied using (5.1.2). The remaining question is therefore to design descriptors that satisfy permutation-invariance and rotation-equivariance. For simplicity we only focus on the case of the potential energy which is invariant with respect to rotations but the arguments can be extended to rotation-equivariance as well.

Apart from the symmetry, a desirable set of properties for descriptors includes smoothness and locality. They should also be complete, meaning they should keep non-equivalent structures distinct [121]. Additionally the computational cost of evaluating these descriptors should be as low as possible. The generic construction proceeds as follows. As mentioned in (5.1.2), the energy is expressed as a sum of atom-centered local energies. For a given atomic environment, the descriptors are designed to depend on at most ν atoms, a parameter referred to as the body order. Then the design of descriptors with body order ν proceeds either via a symmetrization first with respect to the rotations and then the permutations or the opposite, as we present in the next subsections. For example among widely known descriptors is the Coulomb matrix $\mathbf{C} \in \mathbb{R}^{M \times M}$ [128] which is defined as:

$$C_{ij} = \begin{cases} 0.5\, Z_i^{2.4} & \text{if } i = j \\ \frac{Z_i Z_j}{\|\mathbf{R}^i - \mathbf{R}^j\|} & \text{if } i \neq j. \end{cases}$$

This descriptor is rotation-invariant. It is not permutation-invariant but can be made so by sorting the rows and columns by norm, or using eigenvalues.

Many descriptors correspond to low body-order terms. Among classical descriptors are the Behler–Parinello functions [21], the smooth overlap of atomic positions (SOAP) [13], Permutation-Invariant Polynomials (PIPs) [32], scattering transform [56], Moment Tensor Potential (MTP) [136], Atomic Cluster Expansion (ACE) [A15]. Note that in some cases, equivariant descriptors are used even for an invariant function, in order to have a richer space of descriptors. These equivariant features are then summed over in the functional form E_θ to obtain an invariant function. For a review on descriptors, see [111].

5.2.3 Rotation-invariance then permutation-invariance

A simple way to obtain rotation-invariant variables is to consider internal coordinates. These are typically bond lengths (two-body), angles (three-body), or torsions (four-body). These

coordinates are trivally rotation-invariant but not permutation-invariant. To symmetrize over permutations, several options are possible. It is possible to average over permutations. This is doable for low body orders but intractable for even moderate body orders. It is possible to sort the values, with the caveat that the result may not be smooth. Histograms can also be considered.

The symmetry functions proposed by Behler and Parinello in [21] have had a large impact, as they made the first successful attempt of explicitly bringing machine-learning ideas to the construction of interatomic potentials for condensed-phase material. The symmetry functions consist of two-body descriptors called radial symmetry functions

$$G_i^1(\mathbf{R}) = \sum_{j \neq i} e^{-\eta \times (r_{ij} - R_s)^2} \times f_c(r_{ij}),$$

where

$$f_c(r) = \begin{cases} 0.5 \times \left[\cos\left(\pi \times \frac{r}{R_c}\right) + 1\right] & \text{for } r \leq R_c \\ 0 & \text{for } r > R_c, \end{cases}$$

is a cutoff function, $R_c > 0$ the cutoff radius, r_{ij} is the distance between atoms i and j in the system, and η and R_s are parameters that control the width and center of the Gaussian, respectively. There are also three-body descriptors called angular symmetry functions:

$$G_i^2(\mathbf{R}) = 2^{(1-\zeta)} \sum_{j,k \neq i} (1 + \lambda \cos(\theta_{ijk}))^\zeta e^{-\eta(r_{ij}^2 + r_{ik}^2 + r_{jk}^2)} f_c(r_{ij}) f_c(r_{ik}) f_c(r_{jk}),$$

where θ_{ijk} is the angle between vectors $\mathbf{R}^i - \mathbf{R}^j$ and $\mathbf{R}^i - \mathbf{R}^k$, η, ζ, and λ (λ is typically -1 or 1) are parameters that control the shape and behavior of the angular term and Gaussian. In the original article, 48 symmetry functions were used in total.

Many other works employ descriptors based on variants of symmetry functions [137, 146]. In these cases, the low body order (typically two- or three-body) allows for an explicit summation over permutations, making the descriptors computationally efficient. However, a major drawback of such low body orders is the lack of completeness. For instance, it is possible to construct configurations that yield identical histograms of pairwise distances but differ in higher-order geometric features, such as bond angles [14]. These configurations are thus non-equivalent, yet they are indistinguishable from the descriptors.

Other descriptors incorporate higher body orders to capture more complex structural information. A notable example is the permutation-invariant polynomials (PIPs), originally developed by Braams and Bowman [32], extended to condensed phase systems in [A27], called atomic-PIPs in this case.

The main idea is to consider as internal coordinates the interatomic distances and generate permutation-invariant polynomials of these variables. For example, the two-body features are polynomials of the distances r_{ij}. In this case, permutation invariance is trivially satisfied as exchanging atoms i and j does not change the value r_{ij}. For three-body, to simplify notation, we denote the atoms i, j, k by 1,2,3. We consider the three interatomic distances between the atoms 1,2, and 3: r_{12}, r_{13}, r_{23}. Permuting the atoms permutes the distances in a different way. For example, permuting 1 and 2 does not change r_{12} but permutes r_{13} and r_{23}. In this case, the 6 permutations of the atoms 1,2,3 generate all possible 6 permutations over the edges r_{12}, r_{13}, r_{23}. Therefore polynomials that satisfy the permutation symmetry are simply polynomials of the three standard invariant polynomials

$$\begin{aligned} P_1 &= r_{12} + r_{13} + r_{23} \\ P_2 &= r_{12}r_{13} + r_{12}r_{23} + r_{13}r_{23} \\ P_3 &= r_{12}r_{13}r_{23}. \end{aligned}$$

For four-body features, finding invariant polynomials gets more involved. Indeed the permutation over the 4 vertices $1, 2, 3, 4$ only generates a subgroup of possible permutations over the 6 edges $(r_{12}, r_{13}, r_{14}, r_{23}, r_{24}, r_{34})$. To obtain all permutation-invariant polynomials one resorts to invariant theory [50]. One possibility is to numerically compute the primary and secondary invariants using a computer algebra system such as MAGMA [31] from which all invariant polynomials with respect to the specific permutation subgroup can be easily obtained. For the four-body case, there exist 6 primary invariant polynomials $P_1, \dots, P_6$ and 6 secondary invariant polynomials $S_1, \dots, S_6$, such that a basis for the space of polynomials satisfying these symmetries is composed of the following polynomials

$$S_i^{\{0,1\}} \prod_{j=1}^{6} P_j^{a_j}.$$

Provided that the primary and secondary invariants can be computed, this method is generic and can, in principle, handle descriptors of arbitrary body order. Unfortunately, in practice, going beyond body order five has proven to be computationally intractable.

These descriptors up to body-order five combined with a linear function in the descriptors gave very satisfactory results on a Tungsten and Silicon database presented in [A27]. This was later extended to molecules [A1] with good success in particular with extrapolation capabilities to large molecules that were not present in the database. A key aspect of the learning process was the chosen regularization, which was done separately at each body order, ensuring that the different body order components of the potential have a reasonable shape, and in particular do not exhibit too many oscillations. A nice aspect is that due to the linearity of the functional, the energy can be decomposed as functions of 1 variable (two-body), 2 variables (three-body), 6 variables (four-body), and 10 variables (five-body). And the 2-body and 3-body functions can easily be plotted, helping the explainability of the obtained results.

There are, however, two main limitations to this method. First, as mentioned above, generating the invariants becomes expensive and even unfeasible for high body order. Second, the cost of evaluating the potential scales with the number of ν-uplets of neighbors (see (5.1.3)), which also explodes when the body order increases.

5.2.4 Permutation-invariance then rotation-invariance

It turns out that for higher body order descriptors, it is computationally more efficient to first symmetrize with respect to permutations and then with respect to rotations.

Density trick

Most of the descriptors in this category are based on considering some density, which reads, for identical atoms,

$$\forall \mathbf{x} \in \mathbb{R}^d, \quad \rho(\mathbf{x}) = \sum_i g(\mathbf{x} - \mathbf{R}_i), \tag{5.2.1}$$

where g is a chosen function. One advantage is that the density is often a smooth function of the atomic coordinates unlike interatomic distances, and one can, moreover, choose the regularity through the function g.

In order to explain the density trick, we focus for now on a one-dimensional example. Forgetting about the rotation invariance for a moment, a basis for invariant polynomials in 1D

for functions of N variables is

$$\begin{aligned} P_1 &= x_1 + x_2 + \ldots + x_N \\ P_2 &= x_1x_2 + \ldots + x_{N-1}x_N \\ &\vdots \\ P_N &= \prod_{i=1}^{N} x_i. \end{aligned}$$

This is for example the basis used in PIPs for body-order 3. Evaluating these polynomials scales as number of terms in the polynomials P_k, which is $\binom{N}{k}$, therefore the scaling is exponential in N. But this is only one possible basis. An alternative basis is

$$\begin{aligned} Q_1 &= x_1 + x_2 + \ldots + x_N \\ Q_2 &= x_1^2 + x_2^2 + \ldots + x_N^2 \\ &\vdots \\ Q_N &= x_1^N + x_2^N + \ldots + x_N^N. \end{aligned}$$

The number of terms in the polynomials Q_k is always N, and we can precompute $x_1, x_1^2, \ldots, x_2, \ldots, x_2^N$. Therefore the computational cost of evaluating the basis is of order $O(N^2)$. Thus we get rid of the exponential scaling in the number of terms so that evaluating the basis Q is much cheaper than evaluating the basis P, while the two bases generate the same space.

The Q basis is related to the density (5.2.1) as it corresponds to considering different functions g taken as $x \mapsto x^k$. In 3D, an additional symmetrization over rotations is required to ensure rotational invariance. A key advantage of this density trick is that the evaluation of the descriptor at a given body-order scales only linearly with the number of neighbors included within the cutoff around an atom. Therefore considering descriptors based on products of densities is numerically more appealing than the PIPs descriptors.

Enforcing rotational symmetry

The considered descriptors in this context are as follows: three-body terms are based on products of two densities $\rho_i(\mathbf{x})\rho_j(\mathbf{x})$ and four-body terms are based on products of three densities $\rho_i(\mathbf{x})\rho_j(\mathbf{x})\rho_k(\mathbf{x})$. These quantities are clearly permutation-invariant. The question is now how to enforce the rotational symmetry.

The main idea is to integrate over the Haar measure of rotations. For example, for three-body terms, this can be written as

$$\int_Q \rho_i(Q\mathbf{x})\rho_j(Q\mathbf{x})dQ,$$

where Q denotes a rotation. This integral is invariant under rotations by construction. The remaining challenge is to evaluate this integral efficiently. To address this we decompose the functions g_i defining the densities on the spherical harmonics basis, or directly take g_i as a spherical harmonics. Indeed the spherical harmonics form an orthonormal basis on the sphere and have well-understood transformation properties under rotations. They enable a tractable and efficient evaluation of these integrals.

Several descriptors are based on such a formulation, starting with the well-known Smooth Overlap of Atomic Orbitals (SOAP) [14]. In this case, the density is taken as

$$\rho_{j\mathbf{R}}(\mathbf{x}) = \sum_i \exp\big(-\alpha|\mathbf{x} - (\mathbf{R}_i - \mathbf{R}_j)|^2\big),$$

i.e. the function g is taken as a Gaussian function with exponent α. Then the density is expanded in a basis containing products of radial functions and spherical harmonics centered at atom j

$$\rho_{j\mathbf{R}}(\mathbf{x}) = \sum_{nlm} c_{nlm}(\mathbf{R}) g_n(|\mathbf{x}|) Y_l^m(\hat{\mathbf{x}}),$$

The power spectrum and bispectrum presented in this article correspond respectively to three and four body-order contributions based on this density choice. A similar construction is presented in the Spectral Neighbor Analysis Potential (SNAP) [142], also called hyperspherical bispectrum descriptor.

Atomic Cluster Expansion

We now present another type of descriptor which generalizes the power spectrum and bispectrum at any body order, called Atomic Cluster Expansion (ACE). A related version of this descriptor is the Moment Tensor Potential [136]. The ACE has been introduced by Drautz in [52], see also [A15].

To simplify the notation, we consider the variables $\mathbf{R}_i := \mathbf{R}_i - \mathbf{R}_j$ with $\mathbf{R}_j$ the center atom. This descriptor takes a density as a sum of Dirac deltas

$$\rho_{\mathbf{R}}(\mathbf{x}) = \sum_i \delta(\mathbf{x} - \mathbf{R}_i).$$

The density is then projected onto a one-particle basis ϕ_{nlm} to generate A_{nlm} coefficients

$$A_{nlm}(\mathbf{R}) := \langle \phi_{nlm}, \rho_{\mathbf{R}} \rangle = \sum_i \phi_{nlm}(\mathbf{R}_i).$$

These quantities are, of course, permutation-invariant, and we consider ν-order correlations (of $\nu + 1$ body order) defined as

$$\prod_{k=1}^{\nu} A_{n_k l_k m_k}(\mathbf{R}).$$

We finally need to symmetrize with respect to rotations. For this we write the one-body basis functions as a product of radial bases g_n and spherical harmonics Y_l^m for the angular part, centered at the origin (where the center atom lies)

$$\phi_{nlm}(\mathbf{R}_i) = \sum_i g_n(R_i) Y_l^m(\hat{R}_i),$$

where R_i is the length of $\mathbf{R}_i$ and $\hat{R}_i$ is the corresponding point on the sphere S. There holds

$$Y_l^m(Q\hat{R}) = \sum_{\mu=-l}^{l} D_{\mu m}^l(Q) Y_l^\mu(\hat{R}), \qquad \forall \hat{R} \in S^2, Q \in \mathrm{SO}(3), \tag{5.2.2}$$

so defining

$$\mathcal{M}_{\boldsymbol{l}} := \{ \boldsymbol{\mu} \in \mathbb{Z}^N \mid -l_\alpha \leq \mu_\alpha \leq l_\alpha \},$$

we obtain defining $Y_{\boldsymbol{l}}^{\boldsymbol{m}} = \prod_{i=1}^N Y_{l_i}^{m_i}$, for any $Q \in SO(3)$,

$$Y_{\boldsymbol{l}}^{\boldsymbol{m}}(Q\hat{R}) = \sum_{\boldsymbol{\mu} \in \mathcal{M}_{\boldsymbol{l}}} D_{\boldsymbol{\mu m}}^{\boldsymbol{l}}(Q) Y_{\boldsymbol{l}}^{\boldsymbol{\mu}}(\hat{R}) \qquad \text{with} \qquad D_{\boldsymbol{\mu m}}^{\boldsymbol{l}}(Q) = \prod_{\alpha=1}^{N} D_{\mu_\alpha m_\alpha}^{l_\alpha}(Q).$$

The matrices D^l appearing above are the Wigner-D matrices. To state our main result, we need two key properties of the Wigner-D matrices. The first one is called addition of angular momenta and decomposes a product of irreducible representations

$$\forall Q \in SO(3), \quad D^l_{\mu_1 m_1}(Q) D^{l_2}_{\mu_2 m_2}(Q) = \sum_{\lambda=|l_1-l_2|}^{l_1+l_2} C^{\lambda(m_1+m_2)}_{l_1 m_1 l_2 m_2} C^{\lambda(\mu_1+\mu_2)}_{l_1 \mu_1 l_2 \mu_2} D^{\lambda}_{(m_1+m_2)(\mu_1+\mu_2)}(Q),$$

where the coefficients $C^{\lambda(\mu_1+\mu_2)}_{l_1\mu_1 l_2\mu_2}$ are Clebsch–Gordan coefficients. Another, more abstract, way to write it, is

$$D^{l_1} \otimes D^{l_2} = \sum_{l=|l_1-l_2|}^{l_1+l_2} D^l,$$

that is the tensor product of two Wigner-D matrices can be decomposed as a sum of Wigner-D matrices with known indices. The second property is that the integral over rotations is as follows

$$\int_{SO(3)} D^l_{\mu m}(Q)\, dQ = \delta_{l0}\delta_{m0}\delta_{\mu 0}.$$

Thus to find invariant functions, we need to find only the terms leading to total angular momentum of $l = 0$. The general case was presented in [151] and we expressed it in a mathematical statement in [A15].

Lemma 5.2.1. *For $N \geq 2$ and $\boldsymbol{l} \in \mathbb{N}^N$, let*

$$\mathcal{L}_{\boldsymbol{l}} = \Big\{ \boldsymbol{L} = (L_2, L_3, \ldots, L_N) \in \mathbb{N}^{N-1} \,\Big|\, |l_1 - l_2| \leq L_2 \leq l_1 + l_2, \quad \forall 3 \leq i \leq N,\ |L_{i-1} - l_i| \leq L_i \leq L_{i-1} + l_i \Big\},$$

then:

(i) Let $\boldsymbol{\mu}, \boldsymbol{m} \in \mathcal{M}_{\boldsymbol{l}}$, then

$$D^{\boldsymbol{l}}_{\boldsymbol{\mu m}}(Q) = [\mathcal{C}_{\boldsymbol{l}}]_{\boldsymbol{\mu},(\boldsymbol{L},M_N)}\ \boldsymbol{D}^{\boldsymbol{l}}(Q)\ [\mathcal{C}_{\boldsymbol{l}}]^T_{\boldsymbol{m},(\boldsymbol{L},M_N)}, \tag{5.2.3}$$

with the generalized Clebsch-Gordan coefficients $\mathcal{C}_{\boldsymbol{l}}$ and the $\boldsymbol{D}^{\boldsymbol{l}}(Q)$ defined as follows:

$$[\mathcal{C}_{\boldsymbol{l}}]_{\boldsymbol{m},(\boldsymbol{L},M_N)} = C^{L_2 M_2}_{l_1 m_1 l_2 m_2} C^{L_3 M_3}_{L_2 M_2 l_3 m_3} \ldots C^{L_N M_N}_{L_{N-1} M_{N-1} l_N m_N}, \quad \text{where} \tag{5.2.4}$$

$$\boldsymbol{L} = (L_2, \ldots, L_N) \in \mathcal{L}_{\boldsymbol{l}}, \quad -L_N \leq M_N \leq L_N, \quad M_i = \sum_{j=1}^{i} m_i;$$

$$\text{and} \qquad \boldsymbol{D}^{\boldsymbol{l}}(Q) = \operatorname{diag}\left\{ D^{L_N}(Q), \quad \boldsymbol{L} = (L_2, L_3, \ldots, L_N) \in \mathcal{L}_{\boldsymbol{l}} \right\}.$$

(ii) Moreover,

$$\int_{SO(3)} \boldsymbol{D}^{\boldsymbol{l}}(Q)\, dQ = \operatorname{diag}\big(\delta_{L_N} \mathbf{1}\big)_{\boldsymbol{L}=(L_2,L_3,\ldots,L_N)\in\mathcal{L}_{\boldsymbol{l}}}. \tag{5.2.5}$$

For example, taking $\boldsymbol{l} = (1, 1, 2)$, there holds $\mathcal{L}_{\boldsymbol{l}} = \{(1,1), (2,0), (2,1), (2,2), (2,3)\}$, and $\boldsymbol{D}^{\boldsymbol{l}}(Q)$ is the block diagonal matrix

$$\boldsymbol{D}^{\boldsymbol{l}}(Q) = \begin{pmatrix} D^1(Q) & 0 & 0 & 0 & 0 \\ 0 & D^0(Q) & 0 & 0 & 0 \\ 0 & 0 & D^1(Q) & 0 & 0 \\ 0 & 0 & 0 & D^2(Q) & 0 \\ 0 & 0 & 0 & 0 & D^3(Q) \end{pmatrix},$$

where the zeros have to be understood as zero matrices of matching size, and the Wigner matrices $D^i(Q)$ are of size $(2i+1)\times(2i+1)$. Moreover,

$$\bar{D}^{(1,1,2)}_{\boldsymbol{\mu m}} = \int_{SO(3)} \boldsymbol{D}^{(1,1,2)}(Q)\, dQ = \begin{pmatrix} 0_{3,3} & 0 & 0 & 0 & 0 \\ 0 & 1_{1,1} & 0 & 0 & 0 \\ 0 & 0 & 0_{3,3} & 0 & 0 \\ 0 & 0 & 0 & 0_{5,5} & 0 \\ 0 & 0 & 0 & 0 & 0_{7,7} \end{pmatrix},$$

where the sizes of the block matrices are only made specific on the diagonal. Then one can easily deduce that $\bar{D}^{(1,1,2)}_{\boldsymbol{\mu m}}$ defined in above is of rank one. The gist of the result is that generalized Clebsch–Gordan coefficients diagonalize the tensor product of Wigner-D matrices, therefore the rank of this tensor is easily evaluated by looking at the block diagonal matrix $\boldsymbol{D^l}$.

At the end, the considered descriptors are particular linear combinations of

$$A_{\boldsymbol{nlm}}(\mathbf{R}) = \prod_{k=1}^{\nu} A_{n_k l_k m_k}(\mathbf{R}),$$

that are rotation-invariant. The basis functions are therefore of the form

$$b_{\boldsymbol{nlL}}(\mathbf{R}) := \sum_{\boldsymbol{\mu}} [\mathcal{C}_{\boldsymbol{l}}]_{\boldsymbol{\mu},(\boldsymbol{L},0)} A_{\boldsymbol{nl\mu}}(\mathbf{R}).$$

Numerically, the energy can be very accurately fitted at a low computational cost. The article [A15] presents the theoretical basis for this work together with a few numerical results on a Silicon and a Tungsten database. One key theoretical result is that we obtain the exact number of rotation-invariant basis functions from Lemma 5.2.1. The linear dependencies coming from the permutation invariance have been taken care of numerically so far. I am currently working on obtaining the exact number of basis functions in this case as well.

ACE has been since used a lot and extended in various ways in particular with nonlinear functionals such as MACE (Message-passing ACE) [15] and Graph ACE [28], as will be described in Section 5.1.2. In [A28], we used ACE extended for equivariant functions to learn Hamiltonians for materials systems.

Approximation theory results on ACE

In [A2], we theoretically studied ACE, and were interested in how the dimensionality of the polynomial space was reduced due to the permutation invariance. From this we obtained approximation results on multiset functions. Compared to other works, such as [74, 145, 86], presenting approximation results for symmetric (and anti-symmetric) functions we focus here on a polynomial approximation.

For simplicity we explain the main aspects of [A2] for smooth symmetric functions $f : [-1,1]^N \to \mathbb{R}$ for which we only consider permutation symmetry. By f being symmetric we mean that

$$f(x_{\sigma 1}, \ldots, x_{\sigma N}) = f(x_1, \ldots, x_N) \qquad \forall \boldsymbol{x} \in [-1,1]^N, \quad \sigma \in S_N. \tag{5.2.6}$$

A general $f \in C([-1,1]^N)$ can be expanded as a Chebyshev series,

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{v}\in\mathbb{N}^N} \hat{f}_{\boldsymbol{v}} T_{\boldsymbol{v}}(\boldsymbol{x}),$$

where $T_{\boldsymbol{v}} = \otimes_{n=1}^N T_{v_n}$ and T_v are the standard Chebyshev polynomials of the first kind. For later reference we define

$$C_{\mathrm{sym}}([-1,1]^N) := \big\{ f \in C([-1,1]^N) \colon f \text{ is symmetric} \big\}.$$

The objective of this work was to incorporate the symmetry in the approximation scheme and see how much can be gained. Due to our assumption of symmetry, all dimensions are equally important so that the total degree approximation is natural: For $D > 0$, we define

$$f_D(\boldsymbol{x}) := \sum_{\boldsymbol{v}:\|\boldsymbol{v}\|_1 \le D} \hat{f}_{\boldsymbol{v}} T_{\boldsymbol{v}}(\boldsymbol{x}), \tag{5.2.7}$$

from which we immediately obtain the exponential approximation error estimate

$$\|f - f_D\|_{L^\infty} \le M\mu^N \rho^{-D}, \tag{5.2.8}$$

provided that f belongs to a Korobov class,

$$\mathcal{K}(M,\mu,\rho) := \big\{ f \text{ s.t. } \textstyle\sum_{\boldsymbol{v}\in\mathbb{N}^N} \rho^{\|\boldsymbol{v}\|_1} |\hat{f}_{\boldsymbol{v}}| \le M\mu^N \big\}. \tag{5.2.9}$$

Although the term ρ^{-D} suggests an excellent approximation rate, the curse of dimensionality still enters through the prefactor μ^N as well as through the cost of evaluating f_D, which scales as

$$\binom{N+D}{D} \sim \begin{cases} D^N/N!, & \text{as } D \to \infty, \\ N^D/D!, & \text{as } N \to \infty, \end{cases} \tag{5.2.10}$$

where $\binom{N+D}{D}$ is the number of terms in (5.2.8). For large N or large D, this seems better than the naive tensor product approximation leading to $\mathcal{O}(D^N)$ terms, but remains expensive in high dimensions. We therefore incorporate the symmetry into the parameterization. This reduces the number of basis functions, since the coefficients of a symmetric function satisfy

$$\hat{f}_{\boldsymbol{v}} = \hat{f}_{\sigma(\boldsymbol{v})} \qquad \forall \sigma \in S_N.$$

Thus, we can obtain a symmetrised representation

$$f_D(\boldsymbol{x}) = \sum_{\boldsymbol{v}\in\mathbb{N}^N_{\rm ord}:\|\boldsymbol{v}\|_1 \le D} c_{\boldsymbol{v}} \ \operatorname{sym} T_{\boldsymbol{v}}(\boldsymbol{x}) \tag{5.2.11}$$

$$\text{where} \qquad \operatorname{sym} T_{\boldsymbol{v}}(\boldsymbol{x}) = \sum_{\sigma\in S_N} T_{\boldsymbol{v}}(\sigma\boldsymbol{x}) = \sum_{\sigma\in S_N} T_{\sigma\boldsymbol{v}}(\boldsymbol{x}) \tag{5.2.12}$$

and $\mathbb{N}^N_{\rm ord}$ denotes the set of all *ordered* N-tuples, i.e., $\boldsymbol{v} \in \mathbb{N}^N_{\rm ord}$ if $v_1 \le v_2 \le \cdots \le v_N$. When $\boldsymbol{v}$ is not strictly ordered then $\sigma\boldsymbol{v} = \boldsymbol{v}$ for some permutations and hence the coefficient $c_{\boldsymbol{v}}$ is different from $\hat{f}_{\boldsymbol{v}}$.

It is immediate to see that $\operatorname{sym} T_{\boldsymbol{v}}$ form a basis of the space of symmetric polynomials, which in turn is dense in $C_{\rm sym}$. In order to not only reduce the number of basis functions but also the computational cost of evaluating the function, we use the ACE as a basis, that is we define

$$A_v(\boldsymbol{x}) := \sum_{n=1}^N T_v(x_n), \qquad v \in \mathbb{N}, \tag{5.2.13}$$

and form the products

$$\boldsymbol{A}_{\boldsymbol{v}}(\boldsymbol{x}) := \prod_{t=1}^N A_{v_t}(\boldsymbol{x}), \qquad \boldsymbol{v} \in \mathbb{N}^N. \tag{5.2.14}$$

The fact that these polynomials form a basis with the same total degree as the tensor basis means that we can expand the target function in this basis, i.e. there exist coefficients $\tilde{c}_{\boldsymbol{v}}$ such that

$$f_D(\boldsymbol{x}) = \sum_{\boldsymbol{v}\in\mathbb{N}^N_{\rm ord}:\|\boldsymbol{v}\|_1 \le D} \tilde{c}_{\boldsymbol{v}} \boldsymbol{A}_{\boldsymbol{v}}(\boldsymbol{x}). \tag{5.2.15}$$

We showed that the computational cost of evaluating (5.2.15) is directly proportional to the the number of terms, or parameters, which we denote by

$$P(N, D) := \#\big\{\boldsymbol{v} \in \mathbb{N}_{\text{ord}}^N : \|\boldsymbol{v}\|_1 \leq D\big\}.$$

A key observation is that the set $\big\{\boldsymbol{v} \in \mathbb{N}_{\text{ord}}^N : \|\boldsymbol{v}\|_1 = D\big\}$ can be interpreted as the set of all integer partitions of D, of length at most N (indices $v = 0$ do not contribute). There exist various bounds for the number of such partitions that incorporate both N and D, such as [124, Theorem 4.9.2], originally presented in [20],

$$P(N, D) \leq \frac{\left(D + \frac{N(N+1)}{2}\right)^N}{(N!)^2} \sim \frac{D^N}{(N!)^2} \qquad \text{as } D \to \infty, \tag{5.2.16}$$

which unsurprisingly suggests that we gain an additional factor $N!$ in the number of parameters and in the computational cost, compared to the total degree approximation which has asymptotic cost $\binom{N+D}{D} \sim \frac{D^N}{N!}$ as $D \to \infty$. Since we are particularly interested in an N-independent bound we instead used a classical result of Hardy and Ramanujan [75].

Lemma 5.2.2. *For any N, D we have*

$$P(N, D) \leq \frac{1}{8\sqrt{3}D} \exp\left(\pi\sqrt{\tfrac{4}{3}D}\right). \tag{5.2.17}$$

The key property of this bound is that it is independent of the domain dimension N. Inserting this bound in (5.2.8), we obtain the following approximation result in terms of the number of parameters $P = P(N, D)$.

Theorem 5.2.3. *Let $f \in C_{\text{sym}}([-1, 1]^N) \cap \mathcal{K}(M, \mu, \rho)$, then there exists a constant $c > 0$ such that for all $D \geq cN$, the symmetric total degree approximation* (5.2.15) *satisfies*

$$\|f - f_D\|_{L^\infty} \leq C \exp\Big(-\alpha[\log P]^2\Big),$$

where $C, \alpha > 0$ are independent of N and D but may depend on M, μ, ρ.

Thus we obtain a super-algebraic convergence rate, which is better than the algebraic rate that would be obtained with a nonsymmetric approximation due to (5.2.10).

We generalized these results for multivariate functions in dimension d, that is $f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ where each coordinate $\boldsymbol{x}_j \in \Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, and f is invariant under permutations, i.e.,

$$f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = f(\boldsymbol{x}_{\sigma 1}, \ldots, \boldsymbol{x}_{\sigma N}) \qquad \forall \sigma \in S_N. \tag{5.2.18}$$

To obtain such approximation results we generalized the proof of Erdös presented in [57] to estimate the number of parameters for a given degree D, proposing a partition function adapted to this case. With this we obtain the following result valid in any dimension d, upon a few assumptions on the basis used that we do not state for simplicity.

Theorem 5.2.4. *Under some assumptions there exist constants $\alpha, M > 0$ such that*

$$\|f - f_D\|_\infty \leq M e^{-\alpha[\log P]^{1+1/d}}.$$

We also obtain a super-algebraic rate in this case which is sharp in the regime $D \lesssim N^{1+1/d}$. Sharper bounds were, moreover, obtained in the case $D \gg N^{1+1/d}$. Finally, the estimation of the number of parameters was used to provide approximation rates for functions defined on multisets. Open questions include incorporating more complex symmetries such as $O(d)$.

5.3 Learning other quantities of interest

Although I have not worked so much is this direction since the work on ACE [A15], the field is evolving very quickly. Therefore I believe it is worth mentioning a few ongoing development which may have a large impact, and are related to the rest of the manuscript, that is the use of machine learning for electronic structure related quantities.

5.3.1 Learning the Hamiltonian

A first natural extension to learning interatomic potentials is to learn quantities directly related to the electronic structure, such as the Hamiltonian matrix, the electronic density, or the density matrix. Often these quantities are intermediate quantities that can be used to accelerate *ab initio* calculations, in a similar spirit as what is presented in Section 4.2. For simplicity, we now focus on the Hamiltonian, which we addressed in [A28]. In principle this Hamiltonian can arise from different models such as tight-binding or DFT. However a representation using localized basis functions such as Gaussian-type orbitals ensures the equivariance of the Hamiltonian. In this case, when the atomic configuration is rotated, the Hamiltonian transforms according to the spherical harmonics which compose the angular part of the basis functions.

In this work we extended ACE to account for equivariant functions in order to learn Hamiltonians for materials systems. Moreover the Hamiltonian is a matrix, and therefore a possibly large number of coefficients have to be learned, which complicates the learning process. In a similar manner, it is possible to learn the electronic density or the density matrix [152].

5.3.2 Learning the wavefunction

In the past few years, a lot of effort has been put in learning the wavefunction directly, using a variational Monte Carlo method. The variational Monte Carlo method consists in minimizing the energy of the wavefunction by sampling to compute high-dimensional integrals. The wavefunction is parametrized using a well-chosen ansatz. Although the variational Monte Carlo method has been proposed a long time ago [101], a renewed interest has arisen recently due to the use of parameterizations based on descriptors.

A key difference between the potential energy surface and the wavefunction is that on top of the symmetry with respect to the permutation of the nuclei position, the wavefunction needs to be antisymmetric with respect to the permutation of the electron positions. This requires to use compatible ansatzes. Most of them are based on determinants, which are antisymmetric possibly with a symmetric prefactor. E.g. the Slater–Jastrow wavefunction [88] writes for electronic coordinates $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$

$$\Psi_\theta(\mathbf{x}) = \exp\left(-J(\mathbf{x})\right) \cdot \Psi_{\text{Slater}}(\mathbf{x}),$$

with a Slater determinant

$$\Psi_\theta(\mathbf{x}_1, \dots, \mathbf{x}_N) = \det[\phi_i(\mathbf{x}_j)],$$

where the one-body functions ϕ_i have still to be parametrized. This parameterization satisfies the required symmetries. However it is in general incomplete, because the nodal surface of the wavefunction is entirely determined by that of the Slater determinant.

A very successful ansatz is the backflow ansatz [60]. It consists in using a generalized Slater determinant

$$\Psi_\theta(\mathbf{x}_1, \dots, \mathbf{x}_N) = \det[\phi_i(\mathbf{x}_j; \{\mathbf{x}_k\}_{k\neq j})]$$

where the orbitals are invariant with respect to permutations within $\{\mathbf{x}_k\}_{k\neq j}$. Hutter showed that a single term of this form suffices to represent a general anti-symmetric function [86]. Of

course, it is possible to combine the Jastrow factor with the backflow ansatz. This seems to lead to very accurate results, in different works, see e.g. the PauliNet [80], the FermiNet [118], or DeepErwin [131]. This approach seems promising at approximating the solution to the Schrödinger equation. In [P2], we compared two of these approaches with tensor-based methods for small molecules, and the ML approaches were most of the time providing the most accurate results.

5.4 Perspectives

Related to this subject, there are a few research directions that I would like to consider.

- **Approximation results for equivariant functions** In Section 5.2.4 we have studied the number of permutation-invariant N-body basis functions with given total degree. It would be interesting to provide analogous results in the case of rotation-equivariant and permutation-invariant functions. For this, we would first need to evaluate or precisely estimate the dimension of the space of rotation-equivariant and permutation-invariant polynomials. This is actually work in progress. Having such bounds, it would be natural to see if they can be used to extend the approximation results presented in [A2] for multiset functions. It would also be interesting to consider similar approximation results for anti-symmetric functions, applicable in the context of wavefunction approximation.

- **Optimal descriptors** Another question would be to study if there exist optimal descriptors, in the sense that they lead to an optimal energy functional on a given set of physically meaningful configurations. This would probably require studying the electronic structure problem and see if those descriptors can be directly derived from the quantum eigenvalue problem. At that point, these descriptors may not be computable in practice, so one would need to develop a strategy to approximate them in an efficient and accurate way. A possible approach would be to formulate the problem in terms of Gromov–Hausdorf distance, which is a distance which is tailored to compare data in different spaces. Here we could e.g. match the descriptors to the electronic density or density matrix. Another way to consider the problem would be to understand how to choose a minimal set of descriptors from which it is possible to reconstruct any atomic configuration.

- **Uncertainty quantification** There are then many questions related to the error of machine-learned interatomic potentials. For example, could one provide error bounds on the energy ? Another question is whether the error at the quantum level, e.g. discretization coming from the DFT calculation, could be put in an interatomic potential uncertainty quantification framework, in order to better estimate the overall error. Having such uncertainty estimate, it would be interesting to use it in order to develop optimized databases, e.g. using greedy algorithms on the configurations, and including new configurations with specific rules coming from the uncertainty quantification.

- **Other symmetries** In principle the ACE framework is generic to any permutation-invariant and equivariance with respect to another group (rigid-body motion in the case of molecules). It has been successfully applied to particle physics in [110]. It would be nice to widen the applications even more and apply this to other Lie groups such as $SU(n)$.

Bibliography

[1] *Cscs (swiss national supercomputing center) annual report 2021.* https://www.cscs.ch/fileadmin/user_upload/contents_publications/annual_reports/AR2021_Final_WEB.pdf.

[2] P. A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, April 2009.

[3] I. Abraham, R. Abraham, M. Bergounioux, and G. Carlier, *Tomographic reconstruction from a few views: a multi-marginal optimal transport approach*, Applied mathematics & optimization, 75 (2017), pp. 55–73.

[4] M. Agueh and G. Carlier, *Barycenters in the Wasserstein Space*, SIAM Journal of Mathematical Analysis, 43 (2011), pp. 904–924.

[5] A. Alfonsi, R. Coyaud, V. Ehrlacher, and D. Lombardi, *Approximation of optimal transport problems with marginal moments constraints*, Mathematics of Computation, 90 (2020), pp. 689–737.

[6] P. C. Alvarez-Esteban, E. del Barrio, J. A. Cuesta-Albertos, and C. Matrán, *A fixed-point approach to barycenters in wasserstein space*, Journal of Mathematical Analysis and Applications, 441 (2016), pp. 744–762.

[7] D. An, S. Y. Cheng, T. Head-Gordon, L. Lin, and J. Lu, *Convergence of stochastic-extended lagrangian molecular dynamics method for polarizable force field simulation*, J. Comput. Phys., 438 (2021), p. 110338.

[8] F. Andrade, G. Peyre, and C. Poon, *Sparsistency for inverse optimal transport*, arXiv:2310.05461, (2023).

[9] M. G. Armentano and R. G. Durán, *Asymptotic lower bounds for eigenvalues by nonconforming finite element methods*, Electronic Transactions on Numerical Analysis, 17 (2004), pp. 93–101.

[10] J. Barnett, C. Farhat, and Y. Maday, *Neural-network-augmented projection-based model order reduction for mitigating the Kolmogorov barrier to reducibility*, Journal of Computational Physics, 492 (2023), p. 112420.

[11] M. Barrault, Y. Maday, N. C. Nguyen, and A. T. Patera, *An 'empirical interpolation'method: application to efficient reduced-basis discretization of partial differential equations*, C. R. Math., 339 (2004), pp. 667–672.

[12] A. Bartók, M. Payne, R. Kondor, and G. Csányi, *Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons*, Physical Review Letters, 104 (2010), p. 136403.

[13] A. P. Bartók and G. Csányi, *Gaussian approximation potentials: A brief tutorial introduction*, International Journal of Quantum Chemistry, 115 (2015), pp. 1051–1057.

[14] A. P. Bartók, R. Kondor, and G. Csányi, *On representing chemical environments*, Physical Review B, 87 (2013), p. 184115.

[15] I. Batatia, D. P. Kovacs, G. Simm, C. Ortner, and G. Csanyi, *MACE: Higher order equivariant message passing neural networks for fast and accurate force fields*, in Advances in Neural Information Processing Systems, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds., vol. 35, Curran Associates, Inc., 2022, pp. 11423–11436.

[16] S. Batzner, A. Musaelian, L. Sun, M. Geiger, et al., *E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials*, Nature Communications, 13 (2022), pp. 1–11.

[17] F. L. Bauer and C. T. Fike, *Norms and exclusion theorems*, Numerische Mathematik, 2 (1960), pp. 137–141.

[18] N. W. Bazley and D. W. Fox, *Lower bounds for eigenvalues of Schrödinger's equation*, Physical Review, 124 (1961), pp. 483–492.

[19] C. Beattie and F. Goerisch, *Methods for computing lower bounds to eigenvalues of self-adjoint operators*, Numerische Mathematik, 72 (1995), pp. 143–172.

[20] A. G. Beged-dov, *Lower and upper bounds for the number of lattice points in a simplex*, SIAM Journal on Applied Mathematics, 22 (1972), pp. 106–108.

[21] J. Behler and M. Parrinello, *Generalized neural-network representation of high-dimensional potential-energy surfaces*, Physical Review Letters, 98 (2007), p. 146401.

[22] C. Bellis and R. Ferrier, *Numerical homogenization by an adaptive fourier spectral method on non-uniform grids using optimal transport*, Computer Methods in Applied Mechanics and Engineering, 419 (2024), p. 116658.

[23] J.-D. Benamou, G. Carlier, and L. Nenna, *A numerical method to solve multi-marginal optimal transport problems with coulomb cost*, in Splitting Methods in Communication, Imaging, Science, and Engineering, Scientific computation, Springer International Publishing, Cham, 2016, pp. 577–601.

[24] A. Bensalah, A. Nouy, and J. Soffo, *Nonlinear manifold approximation using compositional polynomial networks*, arXiv:2502.05088, (2025).

[25] R. Bhatia, *Positive Definite Matrices*, Princeton University Press, 2009.

[26] F. Bigi, M. F. Langer, and M. Ceriotti, *The dark side of the forces: assessing non-conservative force models for atomistic machine learning*, in Forty-second International Conference on Machine Learning, 2025.

[27] P. Binev, A. Cohen, W. Dahmen, R. DeVore, G. Petrova, and P. Wojtaszczyk, *Convergence rates for greedy algorithms in reduced basis methods*, SIAM Journal on Mathematical Analysis, 43 (2011), pp. 1457–1472.

[28] A. Bochkarev, Y. Lysogorskiy, and R. Drautz, *Graph atomic cluster expansion for semilocal interactions beyond equivariant message passing*, Physical Review X., 14 (2024), p. 021036.

[29] A. Bonito, A. Cohen, R. DeVore, D. Guignard, P. Jantsch, and G. Petrova, *Nonlinear methods for model reduction*, ESAIM: Mathematical Modelling and Numerical Analysis, 55 (2021), pp. 507–531.

[30] M. Born and R. Oppenheimer, *Zur quantentheorie der molekeln*, Ann. Phys., 389 (1927), pp. 457–484.

[31] W. Bosma, J. Cannon, and C. Playoust, *The magma algebra system I: The user language*, Journal of Symbolic Computation, 24 (1997), pp. 235–265.

[32] B. J. Braams and J. M. Bowman, *Permutationally invariant potential energy surfaces in high dimensionality*, International Reviews in Physical Chemistry, 28 (2009), pp. 577–606.

[33] P. Bringmann, M. Feischl, A. Miraçi, D. Praetorius, and J. Streitberger, *On full linear convergence and optimal complexity of adaptive FEM with inexact solver*, Computers and Mathematics with Applications, 180 (2025), pp. 102–129.

[34] G. Caloz and J. Rappaz, *Numerical analysis for nonlinear and bifurcation problems*, Handbook of Numerical Analysis, 5 (1997), pp. 487–637.

[35] E. Cancès, R. Chakir, L. He, and Y. Maday, *Two-grid methods for a class of nonlinear elliptic eigenvalue problems*, IMA Journal of Numerical Analysis, (2017), p. drw053.

[36] E. Cancès, R. Chakir, and Y. Maday, *Numerical analysis of the planewave discretization of some orbital-free and Kohn-Sham models*, ESAIM: Mathematical Modelling and Numerical Analysis, 46 (2012), pp. 341–388.

[37] É. Cancès, V. Ehrlacher, D. Gontier, A. Levitt, and D. Lombardi, *Numerical quadrature in the brillouin zone for periodic schrödinger operators*, Numerische Mathematik, 144 (2020), pp. 479–526.

[38] E. Cancès, G. Kemlin, and A. Levitt, *Convergence analysis of direct minimization and self-consistent iterations*, SIAM Journal on Matrix Analysis and Applications, 42 (2021), pp. 243–274.

[39] C. Carstensen and J. Gedicke, *Guaranteed lower bounds for eigenvalues*, Mathematics of Computation, 83 (2014), pp. 2605–2629.

[40] H. Chen, X. Dai, X. Gong, L. He, and A. Zhou, *Adaptive finite element approximations for kohn–sham models*, SIAM Multiscale Modeling and Simulation, 12 (2014), pp. 1828–1869.

[41] Y. Chen, T. T. Georgiou, and A. Tannenbaum, *Optimal transport for gaussian mixture models*, IEEE Access, 7 (2018), pp. 6269–6278.

[42] K. Cheng, S. Aeron, M. C. Hughes, and E. L. Miller, *Dynamical Wasserstein barycenters for time-series modeling*, Advances in Neural Information Processing Systems, 34 (2021), pp. 27991–28003.

[43] A. Cohen and R. DeVore, *Approximation of high-dimensional parametric PDEs **, Acta Numerica, 24 (2015), pp. 1–159.

[44] A. Cohen, C. Farhat, Y. Maday, and A. Somacal, *Nonlinear compressive reduced basis approximation for PDE's*, Comptes Rendus Mécanique, 351 (2023), pp. 1–18.

[45] S. Cole, M. Eckstein, S. Friedland, and K. Życzkowski, *On quantum optimal transport*, Mathematical Physics, Analysis and Geometry, 26 (2023).

[46] C. Cotar, G. Friesecke, and C. Klüppelberg, *Density functional theory and optimal transportation with coulomb cost*, Communications on Pure and Applied Mathematics, 66 (2013), pp. 548–599.

[47] X. Dai, L. He, and A. Zhou, *Convergence and quasi-optimal complexity of adaptive finite element computations for multiple eigenvalues*, IMA Journal of Numerical Analysis, 35 (2015), pp. 1934–1977.

[48] X. Dai, J. Xu, and A. Zhou, *Convergence and optimal complexity of adaptive finite element eigenvalue computations*, Numerische Mathematik, 110 (2008), pp. 313–355.

[49] J. Delon and A. Desolneux, *A Wasserstein-Type distance in the space of gaussian mixture models*, SIAM Journal of Imaging Sciences, 13 (2020), pp. 936–970.

[50] H. Derksen and G. Kemper, *Computational Invariant Theory*, Springer, December 2015.

[51] R. A. DeVore, *Nonlinear approximation*, Acta Numerica, 7 (1998), pp. 51–150.

[52] R. Drautz, *Atomic cluster expansion for accurate and transferable interatomic potentials*, Physical Review B Condens. Matter, 99 (2019), p. 014104.

[53] R. G. Durán, C. Padra, and R. Rodríguez, *A posteriori error estimates for the finite element approximation of eigenvalue problems*, Mathematical Models and Methods in Applied Sciences, 13 (2003), pp. 1219–1229.

[54] W. Dörfler, *A convergent adaptive algorithm for {P}oisson's equation*, SIAM Journal on Numerical Analysis, 33 (1996), pp. 1106–1124.

[55] V. Ehrlacher, D. Lombardi, O. Mula, and F.-X. Vialard, *Nonlinear model reduction on metric spaces. application to one-dimensional conservative PDEs in wasserstein spaces*, ESAIM: Mathematical Modelling and Numerical Analysis, 54 (2020), pp. 2159–2197.

[56] M. Eickenberg, G. Exarchakis, M. Hirn, S. Mallat, and L. Thiry, *Solid harmonic wavelet scattering for predictions of molecule properties*, Journal of Chemical Physics, 148 (2018), p. 241732.

[57] P. Erdos, *On an elementary proof of some asymptotic formulas in the theory of partitions*, Annals of Mathematics, 43 (1942), pp. 437–450.

[58] A. Ern and M. Vohralík, *Polynomial-degree-robust a posteriori estimates in a unified setting for conforming, nonconforming, discontinuous Galerkin, and mixed discretizations*, SIAM Journal on Numerical Analysis, 53 (2015), pp. 1058–1081.

[59] H. Feshbach, *Unified theory of nuclear reactions*, Reviews of Modern Physics, 36 (1958), pp. 1076–1078.

[60] R. P. Feynman and M. Cohen, *Energy spectrum of the excitations in liquid helium*, Physical Review, 102 (1956), p. 1189.

[61] G. E. Forsythe, *Asymptotic lower bounds for the fundamental frequency of convex membranes*, Pacific Journal of Mathematics, 5 (1955), pp. 691–702.

[62] D. W. Fox and W. C. Rheinboldt, *Computational methods for determining lower bounds for eigenvalues of operators in Hilbert space*, SIAM Review, 8 (1966), pp. 427–462.

[63] G. Friesecke, *Optimal Transport: A Comprehensive Introduction to Modeling, Analysis, Simulation, Applications*, Society for Industrial and Applied Mathematics, Jan. 2024.

[64] G. Friesecke and M. Penka, *The GenCol algorithm for high-dimensional optimal transport: General formulation and application to barycenters and wasserstein splines*, SIAM Journal on Mathematics of Data Science, 5 (2023), pp. 899–919.

[65] W. Gangbo and A. Święch, *Optimal maps for the multidimensional Monge-Kantorovich problem*, Communications on Pure and Applied Mathematics, 51 (1998), pp. 23–45.

[66] S. Giani, L. Grubišić, A. Międlar, and J. S. Ovall, *Robust estimates for hp-adaptive approximations of non-self-adjoint eigenvalue problems*, Numerische Mathematik, 133 (2016), p. 471–495.

[67] F. Goerisch and Z. Q. He, *The determination of guaranteed bounds to eigenvalues with the use of variational methods I*, in Computer arithmetic and self-validating numerical methods (Basel, 1989), vol. 7, Academic Press, Boston, MA, 1990, pp. 137–153.

[68] F. Golse, C. Mouhot, and T. Paul, *On the mean field and classical limits of quantum mechanics*, Communications in Mathematical Physics, 343 (2016), pp. 165–205.

[69] D. Gontier and S. Lahbabi, *Convergence rates of supercell calculations in the reduced hartree- fock model*, ESAIM: Mathematical Modelling and Numerical Analysis, 50 (2016), pp. 1403–1424.

[70] M. Griesemer and D. Hasler, *On the smooth Feshbach–Schur map*, Journal of Functional Analysis, 254 (2008), pp. 2329–2335.

[71] L. Grubišić and J. S. Ovall, *On estimators for eigenvalue/eigenvector approximations*, Mathematics of Computation, 78 (2009), pp. 739–770.

[72] D. Guignard and O. Mula, *Tree-Based Nonlinear Reduced Modeling*, Springer Nature Switzerland, Cham, 2024, pp. 267–298.

[73] S. J. Gustafson and I. M. Sigal, *Mathematical Concepts of Quantum Mechanics*, Springer Science & Business Media, September 2011.

[74] J. Han, Y. Li, L. Lin, J. Lu, J. Zhang, and L. Zhang, *Universal approximation of symmetric and anti-symmetric functions*, arXiv:1912.01765, (2019).

[75] G. H. Hardy and S. Ramanujan, *Asymptotic formulaae in combinatory analysis*, Proceedings of the London Mathematical Society, (1918), pp. 75–115.

[76] D. R. Hartree, *The wave mechanics of an atom with a non-Coulomb central field. Part I. Theory and methods*, Mathematical Proceedings of the Cambridge Philosophical Society, 24 (1928), p. 89.

[77] T. Helgaker, P. Jorgensen, and J. Olsen, *Molecular electronic-structure theory*, John Wiley & Sons, 2014.

[78] P. Henning, A. Målqvist, and D. Peterseim, *Two-level discretization techniques for ground state computations of Bose-Einstein condensates*, SIAM Journal on Numerical Analysis, 52 (2014), pp. 1525–1550.

[79] M. F. Herbst, A. Levitt, and E. Cancès, *DFTK: A julian approach for simulating electrons in solids*, JuliaCon Proceedings, 3 (2021), p. 69.

[80] J. Hermann, Z. Schätzle, and F. Noé, *Deep-neural-network solution of the electronic schrödinger equation*, Nat. Chem., 12 (2020), pp. 891–897.

[81] J. Hesthaven, G. Rozza, and B. Stamm, *Certified Reduced Basis Methods for Parametrized Partial Differential Equations*, SpringerBriefs in Mathematics, Springer, 2016.

[82] J. S. Hesthaven and S. Ubbiali, *Non-intrusive reduced order modeling of nonlinear problems using neural networks*, Journal of Computational Physics, 363 (2018), pp. 55–78.

[83] V. Heuveline and R. Rannacher, *A posteriori error control for finite approximations of elliptic eigenvalue problems*, Advances in Computational Mathematics, 15 (2001), pp. 107–138.

[84] J. Hu, Y. Huang, and Q. Lin, *Lower bounds for eigenvalues of elliptic operators: by nonconforming finite element methods*, Journal of Scientific Computing, 61 (2014), pp. 196–221.

[85] J. Hu, Y. Huang, and Q. Shen, *The lower/upper bound property of approximate eigenvalues by nonconforming finite element methods for elliptic operators*, Journal of Scientific Computing, 58 (2014), pp. 574–591.

[86] M. Hutter, *On representing (anti)symmetric functions*, arXiv:2007.15298, (2020).

[87] A. Iollo and T. Taddei, *Mapping of coherent structures in parameterized flows by learning optimal transportation with gaussian models*, Journal of Computational Physics, 471 (2022), p. 111671.

[88] R. Jastrow, *Many-body problem with strong forces*, Physical Review, 98 (1955), pp. 1479–1484.

[89] L. V. Kantorovich, *Functional analysis and applied mathematics*, Russian Mathematical Surveys, 3 (1948), pp. 89–185.

[90] T. Kato, *On the upper and lower bounds of eigenvalues*, Journal of the Physical Society of Japan, 4 (1949), pp. 334–339.

[91] T. Kato, *Perturbation theory for linear operators*, Springer Berlin Heidelberg, 1976.

[92] D. Kingma and J. Ba, *Adam: A method for stochastic optimization*, International Conference on Learning Representations, (2014).

[93] A. V. Knyazev, *Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method*, SIAM Journal on Scientific Computing, 23 (2001), pp. 517–541.

[94] W. Kohn and L. J. Sham, *Self-consistent equations including exchange and correlation effects*, Physical Review, 140 (1965), pp. A1133–A1138.

[95] J. R. Kuttler and V. G. Sigillito, *Bounding eigenvalues of elliptic operators*, SIAM Journal on Mathematical Analysis, 9 (1978), pp. 768–778.

[96] J. R. Kuttler and V. G. Sigillito, *Estimating eigenvalues with a posteriori/a priori inequalities*, vol. 135 of Research Notes in Mathematics, Pitman Advanced Publishing Program, Boston, MA, 1985.

[97] Y. A. Kuznetsov and S. I. Repin, *Guaranteed lower bounds of the smallest eigenvalues of elliptic differential operators*, Numerische Mathematik, 21 (2013), pp. 135–156.

[98] M. G. Larson, *A posteriori and a priori error analysis for finite element approximations of self-adjoint elliptic eigenvalue problems*, SIAM Journal on Numerical Analysis, 38 (2000), pp. 608–625.

[99] K. Lee and K. T. Carlberg, *Model reduction of dynamical systems on nonlinear manifolds using deep convolutional autoencoders*, Journal of Computational Physics, 404 (2020), p. 108973.

[100] J. E. Lennard-Jones, *Cohesion*, Proceedings of the Physical Society, 43 (1931), p. 461.

[101] W. A. Lester, S. M. Rothstein, and S. Tanaka, *Recent Advances in Quantum Monte Carlo Methods - Part II*, World Scientific, 2002.

[102] X. Liu and S. Oishi, *Verified eigenvalue evaluation for the Laplacian over polygonal domains of arbitrary shape*, SIAM Journal on Numerical Analysis, 51 (2013), pp. 1634–1654.

[103] Y. Maday and A. T. Patera, *Numerical analysis of a posteriori finite element bounds for linear functional outputs*, Mathematical Models and Methods in Applied Sciences, 10 (2000), pp. 785–799.

[104] Y. Maday and G. Turinici, *Error bars and quadratically convergent methods for the numerical simulation of the Hartree-Fock equations*, Numerische Mathematik, 94 (2003), pp. 739–770.

[105] 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 (2012), pp. 1419–1475.

[106] R. J. McCann, *A convexity principle for interacting gases*, Advances in Mathematics, 128 (1997), pp. 153–179.

[107] A. Miraçi, J. Papež, and M. Vohralík, *A-posteriori-steered p-robust multigrid with optimal step-sizes and adaptive number of smoothing steps*, SIAM Journal on Scientific Computing, 43 (2021), pp. S117–S145.

[108] C. B. Moler and L. E. Payne, *Bounds for eigenvalues and eigenvectors of symmetric operators*, SIAM Journal on Numerical Analysis, 5 (1968), pp. 64–70.

[109] G. Monge, *Mémoire sur la théorie des déblais et des remblais*, Mémoires de mathématique et de physique, presentés à l'Académie royale des sciences, (1781).

[110] J. M. Munoz, I. Batatia, and C. Ortner, *Boost invariant polynomials for efficient jet tagging*, Machine Learning: Science and Technology, 3 (2022), p. 04LT05.

[111] F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti, *Physics-inspired structural representations for molecules and materials*, Chemical Reviews Journal, 121 (2021), pp. 9759–9815.

[112] M. Nakao, M. Plum, and Y. Watanabe, *Numerical Verification Methods and Computer-Assisted Proofs for Partial Differential Equations*, Springer, Singapore, 2019.

[113] A. M. N. Niklasson, *Extended Born-Oppenheimer Molecular Dynamics*, Physical Review Letters, 100 (2008), p. 123004.

[114] M. Ohlberger and S. Rave, *Reduced basis methods: Success, limitations and future challenges*, arXiv:1511.02021, (2015).

[115] J. M. Ortega, *The Newton-Kantorovich Theorem*, The American Mathematical Monthly, 75 (1968), p. 658.

[116] E. Perlt, ed., *Basis Sets in Computational Chemistry*, Springer, Cham, 2021.

[117] G. Peyré and M. Cuturi, *Computational optimal transport: With applications to data science*, Foundations and Trends® in Machine Learning, 11 (2019), pp. 355–607.

[118] D. Pfau, J. Spencer, A. Matthews, and W. Foulkes, *Ab initio solution of the many-electron Schrödinger equation with deep neural networks*, Physical Review Research, 2 (2020).

[119] D. H. Pham, *Bases mixtes ondelettes-gaussiennes pour le calcul de structures électroniques*, PhD thesis, Grenoble Alpes, 2017.

[120] M. Plum, *Guaranteed numerical bounds for eigenvalues*, in Spectral theory and computational methods of Sturm-Liouville problems (Knoxville, TN, 1996), vol. 191, Dekker, New York, 1997, pp. 313–332.

[121] S. N. Pozdnyakov, M. J. Willatt, A. P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti, *Incompleteness of atomic structure representations*, Physical Review Letters, 125 (2020), p. 166001.

[122] X. Quan and H. Chen, *A finite element configuration interaction method for Wigner localization*, Journal of Computational Physics, 489 (2023), p. 112251.

[123] J. Rabin, G. Peyré, J. Delon, and M. Bernot, *Wasserstein barycenter and its application to texture mixing*, in Scale Space and Variational Methods in Computer Vision: Third International Conference, May 29–June 2, 2011, Revised Selected Papers 3, Springer, 2012, pp. 435–446.

[124] J. L. Ramírez Alfonsín and J. . Ramírez, *The Diophantine Frobenius Problem*, OUP Oxford, December 2005.

[125] R. Rannacher, A. Westenberger, and W. Wollner, *Adaptive finite element solution of eigenvalue problems: balancing of discretization and iteration error*, Numerische Mathematik, 18 (2010), pp. 303–327.

[126] M. Reed and B. Simon, *Methods of modern mathematical physics IV: Analysis of operators*, Academic Press, New York, 1978.

[127] C. Roothaan, *New developments in molecular orbital theory*, Reviews of Modern Physics, 23 (1951), pp. 69–89.

[128] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, *Fast and accurate modeling of molecular atomization energies with machine learning*, Physical Review Letters, 108 (2012), p. 058301.

[129] Y. Saad, *Numerical methods for large eigenvalue problems*, Manchester University Press, Manchester, 1992.

[130] F. Santambrogio, *Optimal Transport for Applied Mathematicians*, Springer International Publishing, 2015.

[131] M. Scherbela, R. Reisenhofer, L. Gerard, P. Marquetand, and P. Grohs, *Solving the electronic Schrödinger equation for multiple nuclear geometries with weight-sharing deep neural networks*, Nature Computational Science, 2 (2022), pp. 331–341.

[132] A. Schmidt, D. Wittwar, and B. Haasdonk, *Rigorous and effective a-posteriori error bounds for nonlinear problems—application to RB methods*, Advances in Computational Mathematics, 46 (2020), p. 32.

[133] J. Schur, *Über potenzreihen, die im innern des einheitskreises beschränkt sind*, Journal für die reine und angewandte Mathematik, 1917 (1917), pp. 205–232.

[134] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, *SchNet - a deep learning architecture for molecules and materials*, Journal of Chemical Physics, 148 (2018), p. 241722.

[135] I. Šebestová and T. Vejchodský, *Two-sided bounds for eigenvalues of differential operators with applications to Friedrichs, Poincaré, trace, and similar constants*, SIAM Journal on Numerical Analysis, 52 (2014), pp. 308–329.

[136] A. Shapeev, *Moment tensor potentials: A class of systematically improvable interatomic potentials*, SIAM Multiscale Modeling and Simulation, 14 (2016), pp. 1153–1173.

[137] J. S. Smith, O. Isayev, and A. E. Roitberg, *ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost*, Chemical Science, 8 (2017), pp. 3192–3203.

[138] G. Still, *Computable bounds for eigenvalues and eigenfunctions of elliptic differential operators*, Numerische Mathematik, 54 (1988), pp. 201–223.

[139] A. J. Stromme, *Wasserstein barycenters: statistics and optimization*, PhD thesis, Massachusetts Institute of Technology, 2020.

[140] E. Tanguy, J. Delon, and N. Gozlan, *Computing barycentres of measures for generic transport costs*, arXiv:2501.04016, (2024).

[141] G. Temple, *The accuracy of Rayleigh's method of calculating the natural frequencies of vibrating systems*, Proceedings of the Royal Society, 211 (1952), pp. 204–224.

[142] A. Thompson, L. Swiler, C. Trott, S. Foiles, and G. Tucker, *Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials*, Journal of Computational Physics, 285 (2015), pp. 316–330.

[143] R. Verfürth, *A posteriori error estimates for nonlinear problems. Finite element discretizations of elliptic equations*, Mathematics of Computation, 62 (1994), pp. 445–475.

[144] C. Villani, *Optimal Transport*, Springer Berlin Heidelberg, 2009.

[145] E. Wagstaff, F. Fuchs, M. Engelcke, M. A. Osborne, and I. Posner, *Universal approximation of functions on sets*, Journal of Machine Learning Research, 23 (2021), pp. 1–56.

[146] H. Wang, L. Zhang, J. Han, and Weinan, *DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics*, Computer Physics Communications, 228 (2018), pp. 178–184.

[147] Y. Wang, *A posteriori error estimation for electronic structure calculations using ab initio methods and its application to reduce calculation costs*, PhD thesis, 2023.

[148] H. F. Weinberger, *Upper and lower bounds for eigenvalues by finite difference methods*, Communications on Pure and Applied Mathematics, 9 (1956), pp. 613–623.

[149] J. Xu and A. Zhou, *A two-grid discretization scheme for eigenvalue problems*, Mathematics of Computation, 70 (1999), pp. 17–26.

[150] Y. Yang, J. Han, H. Bi, and Y. Yu, *The lower/upper bound property of the Crouzeix–Raviart element eigenvalues on adaptive meshes*, Journal of Scientific Computing, 62 (2015), pp. 284–299.

[151] A. P. Yutsis, I. B. Levinson, and V. V. Vanagas, *Mathematical apparatus of the theory of angular momentum*, Israel Program for Scientific Translations, (1962).

[152] L. Zhang, P. Mazzeo, M. Nottoli, E. Cignoni, L. Cupellini, and B. Stamm, *A symmetry-preserving and transferable representation for learning the Kohn-Sham density matrix*, arXiv:2503.08400, (2025).