HAL
open science

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

Eric Cancès, Geneviève Dusson, Gaspard Kemlin, Antoine Levitt

To cite this version:

Eric Cancès, Geneviève Dusson, Gaspard Kemlin, Antoine Levitt. Practical error bounds for properties in plane-wave electronic structure calculations. SIAM Journal on Scientific Computing, 2022, 44 (5), 10.1137/21M1456224 . hal-03408321v2

HAL Id: hal-03408321

https://inria.hal.science/hal-03408321v2

Submitted on 18 Jun 2022

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

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

PRACTICAL ERROR BOUNDS FOR PROPERTIES IN PLANE-WAVE ELECTRONIC STRUCTURE CALCULATIONS

ERIC CANCÈS[1,2], GENEVIÈVE DUSSON[3], GASPARD KEMLIN[1,2], AND ANTOINE LEVITT[1,2]

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

1. Introduction

This article focuses on providing practical error estimates for numerical approximations of electronic structure calculations. Such computations are key in many domains, as they allow for the simulation of systems at the atomic and molecular scale. They are particularly useful in the fields of chemistry, materials science, condensed matter physics, or molecular biology. Among the many electronic structure models available, Kohn–Sham (KS) Density Functional Theory (DFT) [21] with semilocal density functionals is one of the most used in practice, as it offers a good compromise between accuracy and computational cost. We will focus on this model in this article. Note that the mathematical formulation of this problem is similar to that of the Hartree–Fock [13] or Gross–Pitaevskii equations [29], so that what we present in the context of DFT can be easily extended to such contexts. We will focus on plane-wave discretizations within the pseudopotential approximation, which are most suited for the study of crystals; some (but not all) of our methodology can be applied in other contexts as well, including the aforementioned Hartree–Fock or Gross–Pitaevskii models, as well as molecules simulated using plane-wave DFT.

In this field, the first and most crucial problem is the determination of the electronic ground state of the system under consideration. Mathematically speaking, this problem is a constrained minimization problem. Writing the first-order optimality conditions of this problem leads to an eigenvalue problem that is nonlinear in the eigenvectors. At the continuous level, the unknown is a subspace of dimension N_{el}, the number of electrons in the system; this subspace can be conveniently described using either the orthogonal projector on it (density matrix formalism) or an orthonormal basis of it (orbital formalism). This problem is well-known in the literature and the interested reader is referred to [7] and the references therein for more information on how it is solved in practice.

Solving this problem numerically requires a number of approximations, so that only an approximation of the exact solution can be computed. Being able to estimate the error between the exact and the approximate solutions is crucial, as this information can be used to reduce the high computational cost of such methods by an optimization of the approximation parameters, and maybe more importantly, to add error bars to quantities of interest (QoI) calculated from the approximate solution. In our context, such QoI are typically the ground-state energy of the system and the forces on the atoms in the system, but may also include any information computable from the Kohn–Sham ground state.

While such error bounds have been developed already some time ago for boundary value problems, e.g. in the context of finite element discretization, using in particular *a posteriori* error estimation [32], the development of bounds in the context of electronic structure is quite recent, and still incomplete. Computable and guaranteed error bounds for linear eigenvalue problems have been proposed in the last decade [2, 3, 4, 8, 22]; we refer to [25, Chapter 10] for a recent monograph on the subject. Specifically for electronic structure calculations, some of us proposed guaranteed error bounds for linear eigenvalue equations [17]. For fully self-consistent (nonlinear)

(1) CERMICS, École des ponts
(2) Inria Paris
(3) Laboratoire de Mathématiques de Besançon, UMR CNRS 6623, Université Bourgogne Franche-Comté
E-mail addresses: eric.cances@enpc.fr, genevieve.dusson@math.cnrs.fr, gaspard.kemlin@enpc.fr, antoine.levitt@inria.fr.

eigenvalue equations, an error bound was proposed for a simplified 1D Gross–Pitaevskii equation in [10]; however the computational cost of evaluating the error bound in this contribution is quite high. So far, no error bound has been proposed for the error estimation of general QoI in electronic structure calculations, in particular for the interatomic forces. This is what we are trying to achieve in this contribution.

In this article, we use a general approach based on a linearization of the Kohn–Sham equations. It is instructive to compare our approach to those used in a general context. Assume we want to find $x \in \mathbb{R}^n$ such that $f(x) = 0$, for some nonlinear function $f : \mathbb{R}^n \to \mathbb{R}^n$ (the residual). Near a solution x_*, we have $f(x) \approx f'(x)(x - x_*)$, and therefore, if $f'(x)$ is invertible, we have the error-residual relationship

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

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

$$A(x) - A(x_*) \approx \nabla A(x) \cdot \left(f'(x)^{-1} f(x) \right). \tag{2}$$

From here, we obtain the simple first estimate

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

where $|\cdot|$ is any chosen norm on $\mathbb{R}^n$, and $\|\cdot\|_{\mathrm{op}}$ is the induced operator norms on $\mathbb{R}^{n\times n}$ (note that $\nabla A(x) \in \mathbb{R}^n$ and $f'(x) \in \mathbb{R}^{n\times n}$). This approximate bound can be turned into a rigorous one using information on the second derivatives of f; see for instance [31].

In extending this approach to Kohn–Sham models, we encounter several difficulties:

- The structure of our problem is not easily formulated as above because of the presence of constraints and degeneracies. We solve this using the geometrical framework of [7] to identify the appropriate analog to the Jacobian $f'(x)$.
- The computation of the Jacobian and its inverse is prohibitively costly. We use iterative strategies to keep this cost manageable.
- Choosing the right norm is not obvious in this context. For problems involving partial differential equations, where f includes partial derivatives, it is natural to consider Sobolev-type norms, with the aim of making f' a bounded operator between the relevant function spaces. We explore different choices and their impacts on the error bounds.
- The operator norm inequalities

  $$\left| \nabla A(x) \cdot \left(f'(x)^{-1} f(x) \right) \right| \leqslant |\nabla A(x)| \left\| f'(x)^{-1} \right\|_{\mathrm{op}} |f(x)|$$

  are very often largely suboptimal, even with appropriate norms. We quantify this on representative examples.
- The structure of the residual $f(x)$ plays an important role. For instance, when x results from a Galerkin approximation to a partial differential equation, $f(x)$ is orthogonal to the approximation space. In the context of plane-wave discretizations, this means the residual only contains high-frequency Fourier components. We demonstrate how this impacts the quality of the above bounds when A represents the interatomic forces, in which case its derivative mostly acts on low-frequency components.

The main result of our work therefore lies in the derivation of an efficient, asymptotically accurate, way of approximating $\nabla A(x) \cdot \left(f'(x)^{-1} f(x) \right)$ using the specific structure of the residual $f(x)$ in a plane-wave discretization, where A represents a component of the interatomic forces of the system. This approximation can then be used either to approach the actual error $A(x) - A(x_*)$ in (2) or to improve $A(x)$ by computing $A(x) - \nabla A(x) \cdot \left(f'(x)^{-1} f(x) \right)$, which is a better approximation of $A(x_*)$. These estimates are a new step towards robust and guaranteed *a posteriori* error estimates for Kohn-Sham models: this paper reflects the process that lead to their derivation, by describing the issues raised when applying natural ideas and how we propose to solve these issues.

Throughout the paper, we will provide numerical tests to illustrate our results. All these tests are performed with the DFTK software [18], a recent Julia package solving the Kohn–Sham equations in the pseudopotential approximation using a plane-wave basis, thus particularly well suited for periodic systems such as crystals [23]. We are mostly interested in three QoI: the ground-state energy, the ground-state density, and the interatomic forces, the latter being computed using the Hellmann–Feynman theorem. We will demonstrate the main points with simulations on a simple system (bulk silicon), then present results for more complicated systems.

We will be interested in this paper only in quantifying the discretization error. However, the general framework we develop can be used also to treat other types of error (such as the ones resulting from the iterative solution of the underlying minimization problem). We only consider insulating systems at zero temperature, and do not consider the error due to finite Brillouin zone sampling [6]; extending the formalism to finite temperature and quantifying the sampling error, particularly for metals, is currently under investigation, see [15] for a first step in this direction.

The outline of this article is as follows. In Section 2, we present the mathematical framework related to the solution of the electronic structure minimization problem, describing in particular objects related to the tangent space of the constraint manifold. In Section 3, we present the Kohn–Sham model and the numerical framework in which our tests will be performed. In Section 4, we propose a first crude bound of the error between the exact and the numerically computed solution based on a linearization argument as well as an operator norm inequality, both for the error on the ground-state density matrix and on the forces. In Section 5, we refine this bound by splitting between low and high frequencies, and using a Schur complement to refine the error bound on the low frequencies. Finally, in Section 6, we provide numerical simulations on more involved materials systems, namely on a gallium arsenide system (GaAs) and a titanium dioxide system (TiO_2), showing that the proposed bounds work well in those cases.

2. Mathematical framework

In this section, we present the models targeted by our study, as well as the elementary tools of differential geometry used to derive and compute the error bounds on the QoI.

2.1. General framework.

The work we present here is valid for a large class of mean-field models including different instances of Kohn–Sham models, the Hartree–Fock model, and the time-independent Gross–Pitaevskii model and its various extensions. To study them in a unified way, we use a mathematical framework similar to the one in [7]. To keep the presentation simple, we will work in finite dimension and consider that the solutions of the problem can be very accurately approximated in a given finite-dimensional space of (high) dimension $\mathcal{N}$, which we identify with $\mathbb{C}^{\mathcal{N}}$. We denote by

$$\langle x, y\rangle \coloneqq \mathrm{Re}\,(x^* y)$$

the ℓ^2 inner product of $\mathbb{C}^{\mathcal{N}}$, seen as a vector space over $\mathbb{R}$. We equip the $\mathbb{R}$-vector space of square Hermitian matrices

$$\mathcal{H} \coloneqq \mathbb{C}^{\mathcal{N}\times\mathcal{N}}_{\mathrm{herm}}$$

with the Frobenius inner product $\langle A, B\rangle_{\mathrm{F}} \coloneqq \mathrm{Re}\,(\mathrm{Tr}(A^*B))$. Note that although it is important in applications to allow for complex orbitals and density matrices, the space of Hermitian matrices is not a vector space over $\mathbb{C}$, and therefore we will always consider vector spaces over $\mathbb{R}$.

The density-matrix formulation of the mean-field model in this reference approximation space reads

$$\min\{E(P),\ P \in \mathcal{M}_{\mathcal{N}}\}, \quad \text{where} \quad \mathcal{M}_{\mathcal{N}} \coloneqq \left\{P \in \mathcal{H} \;\middle|\; P^2 = P,\ \mathrm{Tr}(P) = N_{\mathrm{el}}\right\} \tag{3}$$

is the manifold of rank-N_{el} orthogonal projectors (density matrices), and $E : \mathcal{H} \to \mathbb{R}$ is a C^2 nonlinear energy functional. The parameter N_{el} (with $1 \leqslant N_{\mathrm{el}} \leqslant \mathcal{N}$) is a fixed integer depending on the physical model, and not on its discretization in a finite-dimensional space. For mean-field electronic structure models, N_{el} is the number of electrons or electron pairs in the system (hence the notation N_{el}); in the standard Gross–Pitaevskii model, $N_{\mathrm{el}} = 1$. The energy functional E is of the form

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

where H_0 is the linear part of the mean-field Hamiltonian, and E_{nl} a nonlinear contribution depending on the considered model (see Section 3 for the expressions in the Kohn–Sham model). For simplicity of presentation we will ignore spin in the formalism, but we will include it in the numerical simulations (see Remark 2). The set $\mathcal{M}_{\mathcal{N}}$ is diffeomorphic to the Grassmann manifold of N_{el}-dimensional complex vector subspaces of $\mathbb{C}^{\mathcal{N}}$.

Problem (3) always has a minimizer since it consists in minimizing a continuous function on a compact set. This minimizer may or may not be unique, depending on the model and/or the physical system under study. We will not elaborate here on this uniqueness issue, and assume for the sake of simplicity that (3) has a unique minimizer, which we denote by P_*.

Besides the ground-state energy $E(P_*)$, we can compute from P_* various other physical quantities of interest (QoI), for instance, the electronic density and the interatomic forces. We denote such a QoI by $A_* = A(P_*)$.

We consider the case when $\mathcal{N}$ is too large for problem (3) to be solved completely in the reference approximation space. To solve problem (3), we therefore consider a finite-dimensional subspace $\mathcal{X}$ of $\mathbb{C}^{\mathcal{N}}$ of dimension N_b and solve instead the variational approximation of (3) in $\mathcal{X}$, namely

$$\min\{E(P),\ P \in \mathcal{M}_{\mathcal{N}},\ \mathrm{Ran}\,(P) \subset \mathcal{X}\}. \tag{4}$$

Our goal is then to estimate the errors $\|A(P) - A_*\|$ on the QoI A, where P is typically the minimizer of (4), given the variational space $\mathcal{X}$, and the norm is specific to the QoI. The latter can be a scalar, e.g. the ground-state energy, or a finite or infinite dimensional vector, e.g. the interactomic forces, or the ground-state density.

2.2. **First-order geometry.** The manifold $\mathcal{M}_{\mathcal{N}}$ is a smooth manifold. Its tangent space $\mathcal{T}_P\mathcal{M}_{\mathcal{N}}$ at $P \in \mathcal{M}_{\mathcal{N}}$ is given by

$$\begin{aligned}\mathcal{T}_P\mathcal{M}_{\mathcal{N}} &= \{X \in \mathcal{H} \mid PX + XP = X,\ \mathrm{Tr}(X) = 0\}\\ &= \left\{X \in \mathcal{H} \;\middle|\; PXP = 0, P^\perp X P^\perp = 0\right\},\end{aligned}$$

where $P^\perp = 1 - P$ is the orthogonal projection on $\mathrm{Ran}\,(P)^\perp$ for the canonical inner product of $\mathbb{C}^{\mathcal{N}}$. The set $\mathcal{T}_P\mathcal{M}_{\mathcal{N}}$ is the set of Hermitian matrices that are off-diagonal in the block decomposition induced by P and $P^\perp$; more explicitly, if $P = U \begin{pmatrix} I_{N_{\mathrm{el}}} & 0 \\ 0 & 0 \end{pmatrix} U^*$ for some unitary $U \in \mathrm{U}(\mathcal{N})$, then

$$\mathcal{T}_P\mathcal{M}_{\mathcal{N}} = \left\{X = U \begin{pmatrix} 0 & Y^* \\ Y & 0 \end{pmatrix} U^*,\ Y \in \mathbb{C}^{(\mathcal{N}-N_{\mathrm{el}})\times N_{\mathrm{el}}}\right\}.$$

The orthogonal projection $\mathbf{\Pi}_P$ on $\mathcal{T}_P\mathcal{M}_{\mathcal{N}}$ for the Frobenius inner product is given by

$$\forall\ X \in \mathcal{H}, \quad \mathbf{\Pi}_P(X) = PXP^\perp + P^\perp X P = [P,[P,X]] \in \mathcal{T}_P\mathcal{M}_{\mathcal{N}}, \tag{5}$$

where $[A,B] \coloneqq AB - BA$ is the commutator of A and B. Linear operators acting on spaces of matrices are sometimes referred to as *super-operators* in the physics literature. Throughout this paper, super-operators will be written in bold fonts.

The mean-field Hamiltonian is the gradient of the energy at a given point P (again for the Frobenius inner product):

$$H(P) \coloneqq \nabla E(P) = H_0 + \nabla E_{\mathrm{nl}}(P).$$

To simplify the notation, we set

$$H_* = H(P_*) = \nabla E(P_*). \tag{6}$$

The first-order optimality condition for (3) is that $\nabla E(P_*)$ is orthogonal to the tangent space $\mathcal{T}_{P_*}\mathcal{M}_{\mathcal{N}}$, which can be written, using (5) and (6), as $\mathbf{\Pi}_{P_*}H(P_*) = [P_*,[P_*,H(P_*)]] = 0$. This corresponds to the residual

$$R(P) = \mathbf{\Pi}_P H(P) = [P,[P,H(P)]]$$

being zero at P_*. The residual function R can be seen as a nonlinear map from $\mathcal{H}$ to itself, and its restriction to $\mathcal{M}_{\mathcal{N}}$ as a vector field on $\mathcal{M}_{\mathcal{N}}$ since for all $P \in \mathcal{M}_{\mathcal{N}}$, $R(P) \in \mathcal{T}_P\mathcal{M}_{\mathcal{N}}$.

2.3. **Second-order geometry.** We introduce the super-operators $\mathbf{\Omega}(P)$ and $\boldsymbol{K}(P)$, defined at $P \in \mathcal{M}_N$ and acting on $\mathcal{H}$. These operators were already introduced in [7, Section 2.2], but we recall here their definitions for completeness. To simplify the notation, we will set $\boldsymbol{K}_* \coloneqq \boldsymbol{K}(P_*)$, $\mathbf{\Omega}_* \coloneqq \mathbf{\Omega}(P_*)$.

The super-operator $\boldsymbol{K}_* \in \mathcal{L}(\mathcal{H})$ is the Hessian of the energy projected onto the tangent space to $\mathcal{M}_{\mathcal{N}}$ at P_*:

$$\boldsymbol{K}_* \coloneqq \mathbf{\Pi}_{P_*}\boldsymbol{\nabla}^2 E(P_*)\mathbf{\Pi}_{P_*} = \mathbf{\Pi}_{P_*}\boldsymbol{\nabla}^2 E_{\mathrm{nl}}(P_*)\mathbf{\Pi}_{P_*}. \tag{7}$$

By construction, $\mathcal{T}_{P_*}\mathcal{M}_{\mathcal{N}}$ is an invariant subspace of $\boldsymbol{K}_*$. Note that $\boldsymbol{K}_* = 0$ for linear eigenvalue problems, *i.e.* when $E_{\mathrm{nl}} = 0$.

The super-operator $\mathbf{\Omega}_* \in \mathcal{L}(\mathcal{H})$ is defined by

$$\forall\ X \in \mathcal{H}, \quad \mathbf{\Omega}_* X = -[P_*,[H_*,X]]. \tag{8}$$

The tangent space $\mathcal{T}_{P_*}\mathcal{M}_{\mathcal{N}}$ is also an invariant subspace of $\mathbf{\Omega}_*$.

It is shown in [7] that the energy of a density matrix $P = P_* + X + O\left(\|X\|_{\mathrm{F}}^2\right) \in \mathcal{M}_{\mathcal{N}}$ with $X \in \mathcal{T}_{P_*}\mathcal{M}_{\mathcal{N}}$ is $E(P) = E(P_*) + \langle X, (\mathbf{\Omega}_* + \boldsymbol{K}_*)X\rangle_{\mathrm{F}} + o\left(\|X\|_{\mathrm{F}}^2\right)$. The restriction of the operator $\mathbf{\Omega}_* + \boldsymbol{K}_*$ to the invariant subspace

$\mathcal{T}_{P_*}\mathcal{M}_\mathcal{N}$ can therefore be identified with the second-order derivative of E on the manifold $\mathcal{M}_\mathcal{N}$. Since P_* is a minimum, it follows that

$$\boldsymbol{\Omega}_* + \boldsymbol{K}_* \geqslant 0 \quad \text{on } \mathcal{T}_{P_*}\mathcal{M}_\mathcal{N}.$$

Note that in general, the second-order derivative of a function defined on a smooth manifold is not an intrinsic object; it depends not only on the tangent structure of the manifold, but also on the chosen affine connection. However, at the critical point P_* of E on the manifold, the contributions to the second derivative due the connection vanish and the second-order derivative becomes intrinsic.

For our purposes, it will be convenient to define this second-order derivative also outside of P_*. Relying on (7)-(8), we define for any $P \in \mathcal{M}_\mathcal{N}$ the super-operators $\boldsymbol{\Omega}(P) \in \mathcal{L}(\mathcal{H})$ and $\boldsymbol{K}(P) \in \mathcal{L}(\mathcal{H})$ through

$$\forall\, X \in \mathcal{H}, \quad \boldsymbol{\Omega}(P)X = -[P,[H(P),X]] \quad \text{and} \quad \boldsymbol{K}(P)X = \boldsymbol{\Pi}_P \boldsymbol{\nabla}^2 E(P)\boldsymbol{\Pi}_P X. \tag{9}$$

Both $\boldsymbol{\Omega}(P)$ and $\boldsymbol{K}(P)$ admit $\mathcal{T}_P\mathcal{M}_\mathcal{N}$ as an invariant subspace and their restrictions to $\mathcal{T}_P\mathcal{M}_\mathcal{N}$ are Hermitian for the Frobenius inner product. The map $\mathcal{M}_\mathcal{N} \ni P \mapsto \boldsymbol{\Omega}(P)+\boldsymbol{K}(P) \in \mathcal{L}(\mathcal{H})$ is smooth and the restriction of $\boldsymbol{\Omega}(P)+\boldsymbol{K}(P)$ to $\mathcal{T}_P\mathcal{M}_\mathcal{N}$ provides a computable approximation of the second-order derivative of $E : \mathcal{M}_\mathcal{N} \to \mathbb{R}$ in a neighborhood of P_* (whatever the choice of the affine connection).

2.4. **Density matrix and orbitals.** The framework we have outlined above is particularly convenient for stating the second-order conditions, but much too expensive computationally as it requires the storage and manipulation of (low-rank) large matrices. In practice, it is more effective to work directly with orbitals, *i.e.* write for any $P \in \mathcal{M}_\mathcal{N}$

$$P = \Phi\Phi^* = \sum_{i=1}^{N_{\mathrm{el}}} |\phi_i\rangle\langle\phi_i| \tag{10}$$

where $\Phi = (\phi_1|\cdots|\phi_{N_{\mathrm{el}}})$ is a collection of N_{el} orbitals $\phi_i \in \mathbb{C}^\mathcal{N}$ satisfying $\Phi^*\Phi = I_{N_{\mathrm{el}}}$ and $\mathrm{Span}\,(\phi_1,\ldots,\phi_{N_{\mathrm{el}}}) = \mathrm{Ran}\,(P)$, and where we used Dirac's bra-ket notation: for $\phi,\psi \in \mathbb{C}^\mathcal{N}$, $\langle\phi,\psi\rangle = \phi^*\psi$ and $|\phi\rangle\langle\psi| = \phi\psi^*$. Problem (3) can be reformulated as

$$\min\left\{E(\Phi\Phi^*), \Phi \in \mathbb{C}^{\mathcal{N}\times N_{\mathrm{el}}}, \Phi^*\Phi = I_{N_{\mathrm{el}}}\right\}.$$

Note that the orbitals are only defined up to a unitary transform: if $U \in \mathrm{U}(N_{\mathrm{el}})$ is a unitary matrix, then $\widetilde{\Phi} := \Phi U$ and Φ give rise to the same density matrix. This means that the minimizers of this minimization problem are never isolated, which creates technical difficulties that are not present in the density matrix formalism.

Let us fix a $\Phi = (\phi_1|\cdots|\phi_{N_{\mathrm{el}}}) \in \mathbb{C}^{\mathcal{N}\times N_{\mathrm{el}}}$ with $\Phi^*\Phi = I_{N_{\mathrm{el}}}$, and consider an element X of the tangent plane $\mathcal{T}_{\Phi\Phi^*}\mathcal{M}_\mathcal{N}$. By completing Φ to an orthogonal basis and writing out X in this basis, it is easy to see that the constraints $X^* = X$, $PXP = 0$, $P^\perp XP^\perp = 0$ imply that X can be put in the form

$$X = \sum_{i=1}^{N_{\mathrm{el}}} |\phi_i\rangle\langle\xi_i| + |\xi_i\rangle\langle\phi_i| = \Phi\Xi^* + \Xi\Phi^* \tag{11}$$

where $\Xi = (\xi_1|\cdots|\xi_{N_{\mathrm{el}}}) \in \mathbb{C}^{\mathcal{N}\times N_{\mathrm{el}}}$ is a set of orbital variations satisfying $\Phi^*\Xi = 0$. Furthermore, under this condition, Ξ is unique, so that (11) establishes a bijection between $\mathcal{T}_{\Phi\Phi^*}\mathcal{M}_\mathcal{N}$ and $\{\Xi \in \mathbb{C}^{\mathcal{N}\times N_{\mathrm{el}}} \mid \Phi^*\Xi = 0\}$. We will therefore treat equivalently elements of the tangent space $\mathcal{T}_P\mathcal{M}_\mathcal{N}$ either in the density matrix representation X or the orbital representation Ξ, writing

$$\Xi \simeq_\Phi X. \tag{12}$$

This orbital representation of P by Φ is more economical computationally, only requiring the storage and manipulation of orbitals $\Phi \in \mathbb{C}^{\mathcal{N}\times N_{\mathrm{el}}}$ satisfying $\Phi^*\Phi = I_{N_{\mathrm{el}}}$. Similarly, the manipulation of objects X in the tangent plane $\mathcal{T}_{\Phi\Phi^*}\mathcal{M}_\mathcal{N}$ is more efficiently done through their orbital variations $\Xi \in \mathbb{C}^{\mathcal{N}\times N_{\mathrm{el}}}$ satisfying $\Phi^*\Xi = 0$.

All operations on density matrices or their variations can indeed be carried out in this orbital representation. For instance, the computation of the energy can be performed efficiently in practice, as explained in Section 3, and the residual at $P = \Phi\Phi^*$ also has a nice representation in terms of orbitals:

$$R(\Phi\Phi^*) \simeq_\Phi H\Phi - \Phi(\Phi^*H\Phi) \quad \text{with } H \text{ evaluated at } \Phi\Phi^*,$$

which is easily recognized as similar to the residual of a linear eigenvalue problem.

Likewise, operators on $\mathcal{T}_{\Phi\Phi^*}\mathcal{M}_\mathcal{N}$ can be identified in this fashion. For instance,

$$\boldsymbol{\Omega}(\Phi\Phi^*)(\Phi\Xi^* + \Xi\Phi^*) \simeq_\Phi P^\perp(H\Xi - \Xi(\Phi^*H\Phi)) \quad \text{with } H \text{ evaluated at } \Phi\Phi^*. \tag{13}$$

The computation of $\boldsymbol{K}$ can be performed similarly:

$$\boldsymbol{K}(\Phi\Phi^*)(\Phi\Xi^* + \Xi\Phi^*) \simeq_\Phi P^\perp(\delta H \phi_i)_{i=1,\dots,N_{\rm el}} \quad \text{with } \delta H = \frac{\mathrm{d}H}{\mathrm{d}P}(\Phi\Xi^* + \Xi\Phi^*). \tag{14}$$

Finally, note that all the numerical results in this article are performed using the orbital formalism.

Remark 1. Note that the condition that $\Phi^*\Xi = 0$ is not necessary for $\Phi\Xi^* + \Xi\Phi^*$ to define an element of $\mathcal{T}_{\Phi\Phi^*}\mathcal{M}_\mathcal{N}$. However, without this gauge condition, Ξ is not unique. This is simply a manifestation at the infinitesimal level of the noninjectivity of the map $\Phi \mapsto \Phi\Phi^*$ between $\{\Phi \in \mathbb{C}^{\mathcal{N}\times N_{\rm el}}, \Phi^*\Phi = I_{N_{\rm el}}\}$ and $\mathcal{M}_\mathcal{N}$. Because of this, the derivative $\Xi \mapsto \Phi\Xi^* + \Xi\Phi^*$ is not injective between the tangent spaces $\{\Xi \in \mathbb{C}^{\mathcal{N}\times N_{\rm el}}, \Phi^*\Xi + \Xi^*\Phi = 0\}$ and $\mathcal{T}_{\Phi\Phi^*}\mathcal{M}_\mathcal{N}$. In more concrete terms, in the example case where $\phi_1, \dots, \phi_{N_{\rm el}}$ are the first $N_{\rm el}$ basis vectors, any element X is of the form $\begin{pmatrix} 0 & Z^* \\ Z & 0 \end{pmatrix}$ which can be written in the form (11) with $\Xi = \begin{pmatrix} 0 \\ Z \end{pmatrix}$. However, such an X can also be written in the form (11) with $\Xi = \begin{pmatrix} A \\ Z \end{pmatrix}$ for any anti-hermitian matrix A. The gauge condition $\Phi^*\Xi = 0$ forces A to be zero, making Ξ unique. In more formal terms, the map $\Phi \mapsto \Phi\Phi^*$ induces a principal bundle structure on the base space $\mathcal{M}_\mathcal{N}$ (the Grassmann manifold) with total space $\{\Phi \in \mathbb{C}^{\mathcal{N}\times N_{\rm el}}, \Phi^*\Phi = I_{N_{\rm el}}\}$ (the Stiefel manifold) and characteristic fiber $\mathrm{U}(N_{\rm el})$. This naturally splits the tangent space $\{\Xi \in \mathbb{C}^{\mathcal{N}\times N_{\rm el}}, \Phi^*\Xi + \Xi^*\Phi = 0\}$ into the *vertical space* $\{\Phi A,\ A \text{ anti-hermitian}\}$, and a complementary *horizontal space*, which we take to be the orthogonal complement, $\{\Xi \in \mathbb{C}^{\mathcal{N}\times N_{\rm el}} \mid \Phi^*\Xi = 0\}$.

The orbital formalism can be used to give a more concrete interpretation of the first-order optimality condition $R(P_*) = 0$. Indeed, this condition can be rewritten as

$$P_* H_* P_*^\perp = 0, \quad P_*^\perp H_* P_* = 0,$$

from which it follows that P_* and $H_* = H(P_*)$ can be jointly diagonalized in an orthonormal basis:

$$H_*\phi_{*n} = \lambda_{*n}\phi_{*n}, \quad \langle \phi_{*m}, \phi_{*n}\rangle = \delta_{mn}, \quad P_* = \sum_{n=1}^{N_{\rm el}} |\phi_{*n}\rangle\langle\phi_{*n}|. \tag{15}$$

In many applications, the orbitals $\phi_{*1}, \dots, \phi_{*N_{\rm el}}$ spanning the range of P_* (see (15)) are those corresponding to the lowest $N_{\rm el}$ eigenvalues of H_*. This is called the *Aufbau* principle in physics and chemistry. This principle is always satisfied in the (unrestricted) Hartree–Fock setting, and most of the times in the Kohn–Sham setting. Under the *Aufbau* principle, we can assume that the λ_n's are ranked in nondecreasing order. The orbitals ϕ_i, $1 \leqslant i \leqslant N_{\rm el}$, are called occupied, and the orbitals ϕ_a, $N_{\rm el} \leqslant a \leqslant \mathcal{N}$, are called virtual (it is customary to label the occupied orbitals by indices i, j, k, l, and virtual orbitals by indices a, b, c, d). The operator $\boldsymbol{\Omega}_*$ can be written explicitly using the tensor basis $\phi_{*m} \otimes \phi_{*n}$. We have indeed

$$\boldsymbol{\Omega}_* = \sum_{i=1}^{N_{\rm el}} \sum_{a=N_{\rm el}+1}^{\mathcal{N}} (\lambda_a - \lambda_i)\left(|\phi_{*i}\otimes\phi_{*a}\rangle\langle\phi_{*i}\otimes\phi_{*a}| + |\phi_{*a}\otimes\phi_{*i}\rangle\langle\phi_{*a}\otimes\phi_{*i}|\right),$$

and it follows that the lowest eigenvalue of the restriction of $\boldsymbol{\Omega}_*$ to $\mathcal{T}_{P_*}\mathcal{M}_\mathcal{N}$ is $\lambda_{N_{\rm el}+1} - \lambda_{N_{\rm el}} \geqslant 0$. The operator $\boldsymbol{\Omega}_*$ is therefore positive on $\mathcal{T}_{P_*}\mathcal{M}_\mathcal{N}$, and coercive if there is an energy gap between the ${N_{\rm el}}^{\rm th}$ and $(N_{\rm el}+1)^{\rm st}$ eigenvalues of H_* (see e.g. [7]).

2.5. **Metrics on the tangent space.** The isomorphism between $X = \Phi\Xi^* + \Xi\Phi^* \in \mathcal{T}_{\Phi\Phi^*}\mathcal{M}_\mathcal{N}$ and the set of orbital variations $\Xi \in \mathbb{C}^{\mathcal{N}\times N_{\rm el}}$ with $\Phi^*\Xi = 0$ is unitary under the Frobenius inner product up to a factor of 2: $\|X\|_{\rm F}^2 = 2\,\|\Xi\|_{\rm F}^2$.

In practice, it is often advantageous to work using different inner products. This is in particular the case for partial differential equations involving unbounded operators, where using Sobolev-type metrics better respects the natural analytic structure of the problem and therefore allows for better bounds, compare e.g. the results of (5.34) and (5.35) on Figure 4 in [4]. To that end, consider a metric on $\mathbb{C}^\mathcal{N}$ given by

$$\langle \xi_1, \xi_2\rangle_T = \langle \xi_1, T\xi_2\rangle.$$

Here T is a coercive Hermitian operator on $\mathbb{C}^\mathcal{N}$ representing the metric; for instance, taking T to be a discretization of the operator $1 - \Delta$ we recover the classical Sobolev H^1 norm. A basic problem is that the projection $P^\perp$ onto the orthogonal complement of $\mathrm{Ran}\,(P)$ does not necessarily commute with T. As a result, there are various

nonequivalent ways to lift this metric to one on the tangent space $\mathcal{T}_{\Phi\Phi^*}\mathcal{M}_{\mathcal{N}}$. We select here the computationally simplest. The operator

$$M = P^{\perp}T^{1/2}P^{\perp}T^{1/2}P^{\perp} \tag{16}$$

is positive definite on the subspace $\mathrm{Ran}\,(P)^{\perp}$ of $\mathbb{C}^{\mathcal{N}}$, and induces a metric $\langle \xi_1, M\xi_2\rangle$ on that space. The point of this formulation is to make it easy to compute $M^{1/2} = P^{\perp}T^{1/2}P^{\perp}$. Note that, since $P^{\perp}$ and T do not commute, $M^{-1/2} \neq P^{\perp}T^{-1/2}P^{\perp}$. However, $P^{\perp}T^{-1/2}P^{\perp}M^{1/2}$ is well-conditioned, so that computing the action of $M^{-1/2}$ on a vector can be performed efficiently by an iterative algorithm involving repeated applications of the operators $T^{1/2}$ and $T^{-1/2}$. The same holds for M^{-1}. Furthermore, practical numerical results are typically not very sensitive to these issues, so that other (nonequivalent) reasonable alternatives to (16) yield similar results.

The metric on $\mathrm{Ran}\,(P)^{\perp}$ immediately induces a metric on $\mathcal{T}_{\Phi\Phi^*}\mathcal{M}_{\mathcal{N}}$ given by, in the orbital representation associated with Φ,

$$\langle \Xi_1, \Xi_2\rangle_{\boldsymbol{M}} = \mathrm{Re}\,\left(\mathrm{Tr}(\Xi_1^* M \Xi_2)\right) = \sum_{i=1}^{N_{\mathrm{el}}} \mathrm{Re}\,\left(\langle \xi_{1,i}, M\xi_{2,i}\rangle\right),$$

for $\Xi_1 = (\xi_{1,i})_{1\leqslant i\leqslant N_{\mathrm{el}}}, \Xi_2 = (\xi_{2,i})_{1\leqslant i\leqslant N_{\mathrm{el}}}$. This defines an operator $\boldsymbol{M}$ on $\mathcal{T}_{\Phi\Phi^*}\mathcal{M}_{\mathcal{N}}$ through the relationship $\boldsymbol{M}X \simeq_{\Phi} (M\xi_i)_{1\leqslant i\leqslant N_{\mathrm{el}}}$ when $X \simeq_{\Phi} (\xi_i)_{1\leqslant i\leqslant N_{\mathrm{el}}}$. Similarly to M, we can compute powers and inverses of $\boldsymbol{M}$ easily.

This formalism has the disadvantage that the same metric is used for every orbital variation. In practice this may not be sensible, as different orbitals can correspond to different energy ranges. Therefore we slightly modify the above formalism by applying a different metric on each individual orbital variation, following standard practice used in preconditioners for plane-wave density functional theory [26]. Introducing a family $(T_1, \ldots, T_{N_{\mathrm{el}}})$ of coercive Hermitian operators on $\mathbb{C}^{N_{\mathrm{el}}}$, we set

$$M_i \coloneqq P^{\perp}T_i^{1/2}P^{\perp}T_i^{1/2}P^{\perp} \quad \text{and} \quad \boldsymbol{M}X \simeq_{\Phi} (M_i\xi_i)_{1\leqslant i\leqslant N_{\mathrm{el}}}. \tag{17}$$

2.6. **Correspondence rules.** As explained above, the density matrices will be preferred for the mathematical analysis while orbitals will be used in practice in the numerical simulations. We summarize below the correspondence between the density matrix and molecular orbital formulations, and the practical way the different operators we introduced are computed. For a given $P \in \mathcal{M}_{\mathcal{N}}$ and $(\phi_i)_{1\leqslant i\leqslant N_{\mathrm{el}}}$ the associated set of occupied orbitals, there holds

$$\begin{aligned}
P \in \mathcal{M}_{\mathcal{N}} \quad &\leftrightarrow \quad \Phi = (\phi_1|\cdots|\phi_{N_{\mathrm{el}}}) \in \mathbb{C}^{\mathcal{N}\times N_{\mathrm{el}}} \text{ s.t. } P = \Phi\Phi^* && \text{(state)},\\
X \in \mathcal{T}_P\mathcal{M}_{\mathcal{N}} \quad &\leftrightarrow \quad \Xi \in \mathbb{C}^{\mathcal{N}\times N_{\mathrm{el}}} \text{ s.t. } \Phi^*\Xi = 0 && \text{(perturbation)},\\
R(P) = [P,[P,H(P)]] \quad &\leftrightarrow \quad (r_1|\cdots|r_{N_{\mathrm{el}}}) = P^{\perp}H(P)\Phi && \text{(residual)},\\
\boldsymbol{\Omega}(P)(X) \quad &\leftrightarrow \quad P^{\perp}(H\Xi - \Xi(\Phi^*H\Phi)) && \text{(see (13))},\\
\boldsymbol{K}(P)(X) \quad &\leftrightarrow \quad P^{\perp}(\delta H\phi_i)_{1\leqslant i\leqslant N_{\mathrm{el}}} && \text{(see (14))},\\
\boldsymbol{M}^s X \quad &\leftrightarrow \quad (M_i^s\xi_i)_{1\leqslant i\leqslant N_{\mathrm{el}}} \text{ for } s = -1, -1/2, 1/2, 1 && \text{(see (17))}.
\end{aligned}$$

3. The periodic Kohn–Sham problem

3.1. **The continuous problem.** We consider an $\mathcal{R}$-periodic system, $\mathcal{R}$ being a Bravais lattice with unit cell Γ and reciprocal lattice $\mathcal{R}^*$ (the set of vectors G such that $G\cdot R \in 2\pi\mathbb{Z}$ for all $R \in \mathcal{R}$). For the sake of simplicity, we present here the formalism for the (artificial) Kohn–Sham model for a finite system of N_{el} electrons on the unit cell Γ with periodic boundary conditions. This is distinct from the more physical periodic Kohn–Sham problem for an infinite crystal with N_{el} electrons by unit cell, which is usually treated by using the supercell approach and Bloch theorem. Practical computations are performed for the latter model using Monkhorst-Pack Brillouin zone sampling [24] (see also [6] for a mathematical analysis of this method). The mathematical framework is very similar, with additional sums over k points.

At the continuous level, a Kohn–Sham state is described by a density matrix γ, a rank-N_{el} orthogonal projector acting on the space $\mathrm{L}^2_{\#}$ of square integrable periodic functions. Ignoring constant terms modeling interactions

between ions (*i.e.* atomic nuclear and frozen core electrons), the Kohn–Sham energy of γ is given by $E^{\mathrm{KS}}(\gamma) = \mathrm{Tr}(h_0\gamma) + E^{\mathrm{Hxc}}(\rho_\gamma)$ (the superscript Hxc stands for Hartree-exchange-correlation), with

$$h_0 = -\frac{1}{2}\Delta + v_{\mathrm{loc}} + v_{\mathrm{nloc}}, \quad E^{\mathrm{Hxc}}(\rho) = \int_\Gamma \left(\frac{1}{2}\rho V_{\mathrm{H}}(\rho)(x) + e_{\mathrm{xc}}(\rho(x))\right) \mathrm{d}x.$$

In the above expressions, ρ_γ is the density associated with the trace-class operator γ (formally $\rho_\gamma(x) = \gamma(x,x)$ where $\gamma(x,x')$ is the integral kernel of γ), $e_{\mathrm{xc}} : \mathbb{R}_+ \to \mathbb{R}$ a given exchange-correlation energy, and $V_{\mathrm{H}}(\rho)$ the Hartree potential, defined as the unique periodic solution with zero mean of the Poisson equation $-\Delta V_{\mathrm{H}}(\rho) = 4\pi\left(\rho - \fint_\Gamma \rho\right)$. In the pseudopotential approximation that we use in our numerical results, v_{loc} is a local potential given by

$$\forall\, x \in \mathbb{R}^3, \quad V_{\mathrm{loc}}(x) := \sum_{R\in\mathcal{R}} \sum_{j=1}^{N_{\mathrm{at}}} v_{\mathrm{loc}}^j\left(x - (X_j + R)\right), \tag{18}$$

and v_{nloc} a nonlocal potential in Kleinmann-Bylander [19] form given by

$$V_{\mathrm{nloc}} := \sum_{R\in\mathcal{R}} \sum_{j=1}^{N_{\mathrm{at}}} \sum_{a,b=1}^{n_{\mathrm{proj},j}} \left|p_a^j(\cdot - (X_j + R))\right\rangle C_{ab}^j \left\langle p_b^j(\cdot - (X_j + R))\right|, \tag{19}$$

where N_{at} is the number of atoms in Γ, the X_j's are the positions of the atoms inside the unit cell Γ, $v_{\mathrm{loc}}^j : \mathbb{R}^3 \to \mathbb{R}$ is a local radial potential, $n_{\mathrm{proj},j}$ denotes the number of projectors for atom j, and the $p_{ab}^j : \mathbb{R}^3 \to \mathbb{C}$ are given smooth functions. We use in particular the Goedecker–Teter–Hutter (GTH) pseudopotentials [12, 14] whose functional forms for the v_{loc}^j and p_{ab}^j are analytic (v_{loc}^j is a radial Gaussian function multiplied by a radial polynomial, and p_{ab}^j is a radial Gaussian function multiplied by a solid spherical harmonics).

The Kohn–Sham Hamiltonian associated to a density matrix γ is given by

$$h_\gamma = h_0 + V_{\mathrm{H}}(\rho_\gamma) + e'_{\mathrm{xc}}(\rho_\gamma),$$

where $V_{\mathrm{H}}(\rho_\gamma)$ and $e'_{\mathrm{xc}}(\rho_\gamma)$ are interpreted as local (multiplication) operators. Similarly, we have

$$D_\gamma^2(E^{\mathrm{Hxc}}(\rho_\gamma)) \cdot Q = V_{\mathrm{H}}(\rho_Q) + e''_{\mathrm{xc}}(\rho_\gamma)\rho_Q.$$

Remark 2 (Spin)**.** The expressions above are given for a system of "spinless electrons" to accomodate the simple geometrical formalism of Section 2. Real systems (and the numerical simulations we perform in the following sections) include spin; in this case, the energy is $E^{\mathrm{KS}}(\gamma) = 2\mathrm{Tr}(h_0\gamma) + E^{\mathrm{Hxc}}(\rho_\gamma)$, where $\rho_\gamma(x) = 2\gamma(x,x)$, $\nabla E^{\mathrm{KS}}(\gamma) = 2(h_0 + V_{\mathrm{H}}(\rho_\gamma) + e'_{\mathrm{xc}}(\rho_\gamma))$ and $D^2(E^{\mathrm{KS}}(\gamma)) \cdot Q = 4V_{\mathrm{H}}(\rho_Q) + 4e''_{\mathrm{xc}}(\rho_\gamma)\rho_Q$.

3.2. **Discretization.** For each vector G of the reciprocal lattice $\mathcal{R}^*$, we denote by e_G the Fourier mode with wave-vector G:

$$\forall\, x \in \mathbb{R}^3, \quad e_G(x) := \frac{1}{\sqrt{|\Gamma|}} \exp\left(\mathrm{i}G \cdot x\right)$$

where $|\Gamma|$ is the Lebesgue measure of the unit cell Γ. The family $(e_G)_{G\in\mathcal{R}^*}$ is an orthonormal basis of $\mathrm{L}^2_\#$, the space of locally square integrable $\mathcal{R}$-periodic functions (and an orthogonal basis of the $\mathcal{R}$-periodic Sobolev space $\mathrm{H}^s_\#$, endowed with its usual inner product, for any $s \in \mathbb{R}$). In the so-called plane-wave discretization methods, the Kohn–Sham model is discretized using the finite-dimensional approximation spaces

$$\mathcal{X}_{E_{\mathrm{cut}}} := \mathrm{Span}\left\{e_G,\ G \in \mathcal{R}^* \;\middle|\; \frac{1}{2}|G|^2 \leqslant E_{\mathrm{cut}}\right\},$$

where $E_{\mathrm{cut}} > 0$ is a given energy cut-off chosen by the user.

The connection with the formalism introduced in Section 2 is the following:

- we choose a large reference energy cut-off $E_{\mathrm{cut,ref}}$ and set

$$\mathcal{N} := \dim(\mathcal{X}_{E_{\mathrm{cut,ref}}}) = \#\left\{G \in \mathcal{R}^* \mid \frac{1}{2}|G|^2 \leqslant E_{\mathrm{cut,ref}}\right\};$$

- we identify $\mathcal{X}_{E_{\mathrm{cut,ref}}}$ with $\mathbb{C}^{\mathcal{N}}$ by labelling the reciprocal lattice vectors from 1 to $\mathcal{N}$ in such a way that for all $1 \leqslant i < j \leqslant \mathcal{N}$, $|G_i| \leqslant |G_j|$;

- the set of rank-$N_{\rm el}$ orthogonal projectors γ on $\mathrm{L}^2_\#$ such that $\mathrm{Ran}\,(\gamma) \subset \mathcal{X}_{E_{\rm cut,ref}}$ can then be identified with the manifold $\mathcal{M}_{\mathcal{N}}$ defined in (3) through the mapping
$$\gamma = \sum_{i,j=1}^{\mathcal{N}} P_{ij} \left|e_{G_i}\right\rangle \left\langle e_{G_j}\right|;$$
- the noninteracting Hamiltonian matrix $H_0 \in \mathbb{C}^{\mathcal{N}\times\mathcal{N}}$ has entries
$$[H_0]_{ij} = \left\langle e_{G_i}\middle|h_0\middle|e_{G_j}\right\rangle_{\mathrm{L}^2_\#},$$
and the nonlinear component of the energy $E_{\rm nl} : \mathcal{H} \to \mathbb{R}$ is any C^2-extension of the function defined on $\mathcal{M}_{\mathcal{N}}$ by
$$E_{\rm nl}(P) = E^{\rm Hxc}(\rho_P) \quad \text{where} \quad \rho_P(x) = |\Gamma|^{-1/2} \sum_{i,j=1}^{\mathcal{N}} P_{ij} e_{G_j - G_i}(x).$$

The entries of the core Hamiltonian matrix can be computed explicitly:
$$[H_0]_{ij} = \frac{|G_i|^2}{2}\delta_{i,j} + [V_{\rm loc}]_{ij} + [V_{\rm nloc}]_{ij}$$
$$\text{with} \quad [V_{\rm loc}]_{ij} = \left\langle e_{G_i}\middle|V_{\rm loc}\middle|e_{G_j}\right\rangle_{\mathrm{L}^2_\#} \quad \text{and} \quad [V_{\rm nloc}]_{ij} = \left\langle e_{G_i}\middle|V_{\rm nloc}\middle|e_{G_j}\right\rangle_{\mathrm{L}^2_\#},$$
where the above inner products can be computed exactly through the Fourier transforms of the $v^j_{\rm loc}$ and p^j_{ab} (known exactly for GTH pseudopotentials). Note also that the density ρ_P can be expanded on a finite number of Fourier modes and can therefore be easily stored in memory. Since the Poisson equation is trivially solvable in the plane-wave basis, this enables the exact computation of the Hartree energy. The exchange-correlation energy however cannot be computed explicitly, and is approximated using numerical quadrature. In all the numerical results, we select the parameters of this numerical quadrature such that it does not affect too much the results, see Remark 5.

3.3. **Forces.** The total ground-state energy depends on the atomic positions $\mathfrak{X} = (X_j)_{1\leqslant j\leqslant N_{\rm at}}$ both explicitly (ion-ion interaction energy and ion-electron interaction potentials $V_{\rm loc}$ and $V_{\rm nloc}$) and through the fact that the solution P_* depends on $\mathfrak{X}$:
$$\mathcal{E}(\mathfrak{X}) = E(\mathfrak{X}, P_*(\mathfrak{X})).$$
The force acting on atom j is defined as $F_j(\mathfrak{X}) = -\nabla_{X_j}\mathcal{E}(\mathfrak{X})$. Because of the Hellman–Feynman theorem, the term involving the derivative of P_* with respect to X_j vanishes [23], and the final result is
$$F_j = -\mathrm{Tr}\left((\nabla_{X_j}(V_{\rm loc} + V_{\rm nloc}))P_*\right). \tag{20}$$
This involves the partial derivatives of the matrix elements of $V_{\rm loc} + V_{\rm nloc}$ with respect to the atomic positions, which can be computed analytically from (18) and (19).

3.4. **Numerical setup.** For all the computations and examples on silicon, we use the DFTK software [18] within the LDA approximation, with Teter 93 exchange-correlation functional [12] and a $2\times2\times2$ k-point grid, and a reference solution computed with $E_{\rm cut,ref} = 125$ Ha, to which we compare results obtained with smaller values of $E_{\rm cut}$. We checked that $E_{\rm cut,ref} = 125$ Ha was a high enough energy cut-off to have fully converged results, up to the accuracy we need to test our numerical methods. We use the usual periodic lattice for the FCC phase of silicon, with lattice constant $a = 10.26$ bohrs, close to the equilibrium configuration. All results are expressed in atomic units: energies are in hartree and forces are in hartree/bohr. Note that the discretization grid of the Brillouin zone is not fine enough to have fully converged results, but is still sufficient to illustrate our points. Note also that the same results are observed for semilocal functionals, such as PBE-GGA [27]. Other functionals, such as meta-GGA and hybrid functionals, are out of the scope of this paper.

The two atoms of silicon inside a cell are placed at first at their equilibrium positions with fractional coordinates $(-\frac{1}{8}, -\frac{1}{8}, -\frac{1}{8})$ and $(\frac{1}{8}, \frac{1}{8}, \frac{1}{8})$, and then the second one is slightly displaced by $\frac{1}{20}(0.24, -0.33, 0.12)$ to get nonzero interatomic forces.

The discretized Kohn–Sham equations are solved by a standard SCF procedure. The main computational bottleneck is the partial diagonalization of the mean-field Hamiltonian at each SCF step. This is done using an iterative eigenvalue solver, which only requires applying mean-field Hamiltonian matrices to a set of $N_{\rm el}$ trial orbitals and simple operations on vectors. In a plane-wave basis set of size N_b, the former operation can be done efficiently

through the use of the fast Fourier transform for a total cost of $O(N_{\rm el}N_b(\log N_b + \sum_j n_{{\rm proj},j}))$. We refer to [23] for more details. The application of the super-operators $\boldsymbol{\Omega}$ and $\boldsymbol{K}$ to a set of $N_{\rm el}$ orbital variations (see (13) and (14)) involves additional linear algebra operations, for an additional cost of $O({N_{\rm el}}^2(N_b + N_{\rm el}))$.

In this setting, the reference values for the energy is $E_* = -7.838$ Ha and the interatomic forces are, in hartree/bohr,

$$F_* = \begin{bmatrix} -0.0656 & 0.0656 \\ 0.0619 & -0.0619 \\ -0.0352 & 0.0352 \end{bmatrix},$$

where the first column are the forces acting on the first atom in each direction, and the second column are the forces acting on the second atom.

4. A first error bound using linearization

Now that the mathematical and numerical frameworks are laid down, we turn to the estimation of the error between the reference solution computed with a large energy cut-off $E_{\rm cut,ref}$ and approximations thereof. We first start by deriving a linearization estimate and illustrating numerically its applicability. We then propose a very coarse bound on the error on the density matrix and the forces, based on the (expensive) evaluation of an operator norm. We will show in the next section how to improve this bound.

4.1. **Linearization in the asymptotic regime.** We assume that P_* is a nondegenerate local minimizer of E in the sense that there exists $\eta > 0$ such that $\boldsymbol{\Omega}_* + \boldsymbol{K}_* \geqslant \eta$ on the tangent space $\mathcal{T}_{P_*}\mathcal{M}_{\mathcal{N}}$. This implies in particular that $\boldsymbol{\Omega}_* + \boldsymbol{K}_*$ is invertible on the invariant subspace $\mathcal{T}_{P_*}\mathcal{M}_{\mathcal{N}}$.

Recall that for any trial density matrix $P \in \mathcal{M}_{\mathcal{N}}$, the residual of the problem is

$$R(P) = \boldsymbol{\Pi}_P H(P) = [P, [P, H(P)]] \in \mathcal{T}_P\mathcal{M}_{\mathcal{N}},$$

so that R defines a smooth vector field on $\mathcal{M}_{\mathcal{N}}$ (a section of the tangent bundle $\mathcal{T}_P\mathcal{M}_{\mathcal{N}}$) which vanishes at P_*. For $P \in \mathcal{M}_{\mathcal{N}}$ in the vicinity of P_*, we have

$$P - P_* = \boldsymbol{\Pi}_{P_*}(P - P_*) + O\left(\|P - P_*\|_{\rm F}^2\right) = \boldsymbol{\Pi}_P(P - P_*) + O\left(\|P - P_*\|_{\rm F}^2\right). \tag{21}$$

It follows from the definitions (7)-(8) of $\boldsymbol{\Omega}_*$ and $\boldsymbol{K}_*$ that $\boldsymbol{\Omega}_* + \boldsymbol{K}_*$ is the Jacobian of the map $P \mapsto R(P)$ at P_*. Therefore, the optimality condition $R(P_*) = 0$ and the above expansions yield, for all $P \in \mathcal{M}_{\mathcal{N}}$ close enough to P_*,

$$\begin{aligned} R(P) &= (\boldsymbol{\Omega}_* + \boldsymbol{K}_*)\boldsymbol{\Pi}_{P_*}(P - P_*) + O\left(\|P - P_*\|_{\rm F}^2\right) \\ &= (\boldsymbol{\Omega}(P) + \boldsymbol{K}(P))\boldsymbol{\Pi}_P(P - P_*) + O\left(\|P - P_*\|_{\rm F}^2\right). \end{aligned} \tag{22}$$

By continuity, $\boldsymbol{\Omega}(P) + \boldsymbol{K}(P) \geqslant \frac{\eta}{2}$ on the tangent space $\mathcal{T}_P\mathcal{M}_{\mathcal{N}}$ for $P \in \mathcal{M}_{\mathcal{N}}$ close enough to P_*, so that the restriction of the super-operator $\boldsymbol{\Omega}(P) + \boldsymbol{K}(P)$ to the invariant subspace $\mathcal{T}_P\mathcal{M}_{\mathcal{N}}$ is self-adjoint and invertible. Using again (21) and the fact that $R(P) \in \mathcal{T}_P\mathcal{M}_{\mathcal{N}}$, we obtain, after inversion of $\boldsymbol{\Omega}(P) + \boldsymbol{K}(P)$ on the tangent space,

$$\boxed{P - P_* = ((\boldsymbol{\Omega}(P) + \boldsymbol{K}(P))|_{\mathcal{T}_P\mathcal{M}_{\mathcal{N}}})^{-1}R(P) + O\left(\|P - P_*\|_{\rm F}^2\right).} \tag{23}$$

This error-residual equation is the analog in our case of the linearization (1), which identifies the super-operator $\boldsymbol{\Omega}(P) + \boldsymbol{K}(P)$ as the fundamental object in our study.

Based on this expansion, we can formulate the Newton algorithm to solve the equation $R(P_*) = 0$:

$$P^{k+1} = \mathfrak{R}_{P^k}\left(P^k - (\boldsymbol{\Omega}^k + \boldsymbol{K}^k)^{-1}R(P^k)\right),$$

where $\boldsymbol{\Omega}^k := \boldsymbol{\Omega}(P^k)|_{\mathcal{T}_{P_k}\mathcal{M}_{\mathcal{N}}}$ and $\boldsymbol{K}^k := \boldsymbol{K}(P^k)|_{\mathcal{T}_{P_k}\mathcal{M}_{\mathcal{N}}}$ and $\mathfrak{R}$ is a suitable retraction on $\mathcal{M}_{\mathcal{N}}$. A possible retraction is given in [7]. This Newton algorithm is expensive in practice, as it requires to solve iteratively a linear system; the cost of a Newton step is comparable to that of a full self-consistent field cycle. It is however a useful theoretical tool, and a starting point for further analysis and approximations.

To check the validity of the linearization (23), we focus on three quantities of interest: the ground state energy, the ground-density density, and the interatomic forces acting on the two atoms in Γ. The reference values E_*, ρ_* and F_* of these QoIs are those obtained with the very large energy cut-off $E_{\rm cut,ref} = 125$ Ha, defining a "fine grid" in

real space via the discrete Fourier transform. For $E_{\rm cut} < E_{\rm cut,ref}$ defining a "coarse grid" in real space, we compute two approximations of the three QoIs:

(1) $E_{\rm SCF}$, $\rho_{\rm SCF}$ and $F_{\rm SCF}$ denote the approximations obtained from the variational solution of the Kohn–Sham problem on the coarse grid;
(2) $E_{\rm Newton}$, $\rho_{\rm Newton}$ and $F_{\rm Newton}$ denote the ones computed from the Kohn–Sham state obtained by one Newton step on the fine grid, starting from the variational solution of the Kohn–Sham problem on the coarse grid: as the SCF is converged on the coarse grid, we perform the Newton step on the fine grid in order to improve the approximation of P_*. That is, if P is the variational solution on the coarse grid, $P - ((\boldsymbol{\Omega}(P) + \boldsymbol{K}(P))|_{\mathcal{T}_P\mathcal{M}_\mathcal{N}})^{-1}R(P)$ is a much better approximation of P_* (for the metrics adapted to the chosen three QoIs).

The errors between these approximations and the reference values are plotted in Figure 1 as functions of $E_{\rm cut}$. The errors on the ground-state density are measured with the $\mathrm{L}^2_\#$ metric, while the errors on the forces are measured with the Euclidean metric on $\mathbb{R}^{3\times 2}$.

For the simple case of a silicon crystal at the LDA level of theory, the linearization works very well, even for very small values of $E_{\rm cut}$'s of the order of 5 Ha. Indeed the Kohn–Sham ground-state obtained by variational approximation on a coarse grid is significantly improved by one Newton step: the errors on the QoIs obtained with the latter are orders of magnitude smaller than the ones obtained with the former.

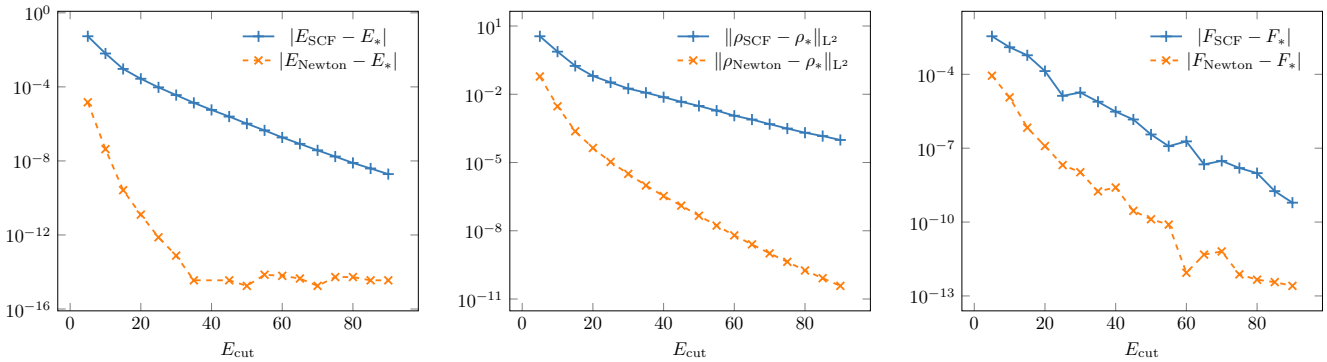

FIGURE 1 – Errors for the ground-state energy (hartree), ground-state density and interatomic forces (hartree/bohr) for Si as a function of $E_{\rm cut}$, for both the variational solution of the Kohn–Sham problem on the coarse grid defined by $E_{\rm cut}$ (solid line) and the post-processed solution obtained with one Newton step on the fine grid (dashed line). This shows that the linearization approximation is excellent, even for energy cut-offs as low as $E_{\rm cut} = 5$ Ha.

4.2. **A simple error bound based on operator norms.** From (23) one can extract an error bound:

$$\begin{aligned} \|P - P_*\| &\approx \|\boldsymbol{\Pi}_P(P - P_*)\| \\ &\leqslant \left\|((\boldsymbol{\Omega}(P) + \boldsymbol{K}(P))|_{\mathcal{T}_P\mathcal{M}_\mathcal{N}})^{-1}\right\|_{\rm op} \|R(P)\| \quad (+\text{ h.o.t.}), \end{aligned} \tag{24}$$

where $\|\cdot\|_{\rm op}$ is the (super-)operator norm associated with the chosen norm $\|\cdot\|$ on $\mathcal{H}$. This bound is not guaranteed, but the results in Figure 1 suggest that it is very close to be guaranteed. Guaranteeing this bound could be done, provided that one could bound the higher-order terms rigorously [31]; this is an interesting prospect, but lies outside the scope of this paper. To test the accuracy of this bound for a specific norm on $\mathcal{H}$, we would need to estimate the corresponding operator norm of the Hermitian operator $((\boldsymbol{\Omega}(P)+\boldsymbol{K}(P))_{\mathcal{T}_P\mathcal{M}_\mathcal{N}})^{-1}$ for all the P's we are considering. In order to lower the computational burden, we consider instead the bound

$$\|P - P_*\| \leqslant \left\|((\boldsymbol{\Omega}_* + \boldsymbol{K}_*)|_{\mathcal{T}_{P_*}\mathcal{M}_\mathcal{N}})^{-1}\right\|_{\rm op} \|R(P)\| \quad (+\text{ h.o.t.}). \tag{25}$$

This enables us to compute the operator norm $\left\|((\boldsymbol{\Omega}_* + \boldsymbol{K}_*)|_{\mathcal{T}_{P_*}\mathcal{M}_\mathcal{N}})^{-1}\right\|_{\rm op}$ only once, instead of computing it for every P. This is of course not accessible in practice, but we use it here for the sake of numerical experiment. Moreover, we can consider that the bounds (24) and (25) are almost equivalent since the results obtained in the previous section show that we are in the linear regime even for the lowest values of $E_{\rm cut}$ used in practice. The operator $(\boldsymbol{\Omega}_* + \boldsymbol{K}_*)|_{\mathcal{T}_{P_*}\mathcal{M}_\mathcal{N}}$ is Hermitian for the Frobenius inner product and, thus, the operator norm

$\left\|((\boldsymbol{\Omega}_* + \boldsymbol{K}_*)_{\mathcal{T}_{P_*}\mathcal{M}_\mathcal{N}})^{-1}\right\|_{\mathrm{op}}$ corresponding to the Frobenius norm on $\mathcal{H}$ is equal to the inverse of the smallest eigenvalue of $(\boldsymbol{\Omega}_* + \boldsymbol{K}_*)|_{\mathcal{T}_{P_*}\mathcal{M}_\mathcal{N}}$. Standard iterative eigenvalue solvers for Hermitian operators can be used to compute this eigenvalue. We use here the LOBPCG algorithm [20].

We can see on Figure 2 (left panel) that when choosing the Frobenius norm on $\mathcal{H}$, the bound (25) leads to very crude error estimates: the error is overestimated by several orders of magnitude, and the bound becomes worse and worse as E_{cut} increases. This issue is well-known in the analysis of partial differential equations, where L^2-type norms are not the natural ones to measure the error on the solution or the residual. Instead, for the Kohn–Sham equations and other second-order elliptic problems, it is more relevant to measure the error $P - P_*$ in H^1-type Sobolev norms (energy norms) and the residual $R(P)$ in H^{-1}-type Sobolev norms (dual norms). The linear operator linking the two quantities (here $(\boldsymbol{\Omega}(P) + \boldsymbol{K}(P))|_{T_P\mathcal{M}_\mathcal{N}}$) is then expected to be a bounded isomorphism from the state error to the residual space for these norms. This suggests adapting the metrics on the tangent space $T_P\mathcal{M}_\mathcal{N}$ in which we measure the error $P - P_*$ (or more precisely the leading term $\boldsymbol{\Pi}_P(P - P_*)$) on the one hand, and the residual $R(P)$ on the other hand. Similar considerations lead to the "kinetic energy preconditioning" used in practical computations [26]. Using the super-operator $\boldsymbol{M}$ on $\mathcal{T}_P\mathcal{M}_\mathcal{N}$ introduced in (17) with T_i the diagonal operator on $\mathbb{C}^\mathcal{N}$ representing the operator $-\frac{1}{2}\Delta + t_i$ where $t_i = \frac{1}{2}\left\|\nabla\phi_i\right\|^2_{\mathrm{L}^2_\#}$ (kinetic energy of the ith orbital), we obtain the bound

$$\left\|\boldsymbol{M}^{1/2}\boldsymbol{\Pi}_P(P - P_*)\right\|_{\mathrm{F}} \leqslant \left\|\boldsymbol{M}^{1/2}((\boldsymbol{\Omega}(P) + \boldsymbol{K}(P))|_{\mathcal{T}_P\mathcal{M}_\mathcal{N}})^{-1}\boldsymbol{M}^{1/2}\right\|_{\mathrm{op}} \left\|\boldsymbol{M}^{-1/2}R(P)\right\|_{\mathrm{F}}. \tag{26}$$

Here also, we lower the computational burden by replacing the first term in the RHS by the asympotically equal quantity $\left\|\boldsymbol{M}_*^{1/2}((\boldsymbol{\Omega}_* + \boldsymbol{K}_*)|_{\mathcal{T}_{P_*}\mathcal{M}_\mathcal{N}})^{-1}\boldsymbol{M}_*^{1/2}\right\|_{\mathrm{op}}$. The results are shown in Figure 2 (central panel). This time, the curves are almost parallel: the gap does not widen as E_{cut} increases. However, the bound is still an overestimate by more than one order of magnitude. This is due to the fact that

$$\left\|\boldsymbol{M}_*^{1/2}((\boldsymbol{\Omega}_* + \boldsymbol{K}_*)|_{\mathcal{T}_{P_*}\mathcal{M}_\mathcal{N}})^{-1}\boldsymbol{M}_*^{1/2}\right\|_{\mathrm{op}} \approx 14.85$$

for this system, while the residual $R(P)$ is supported only on high-frequency Fourier modes, on which the operator $\boldsymbol{M}^{1/2}((\boldsymbol{\Omega}(P) + \boldsymbol{K}(P))|_{\mathcal{T}_P\mathcal{M}_\mathcal{N}})^{-1}\boldsymbol{M}^{1/2}$ is close to identity. The latter statement is supported by Proposition 1 in the appendix (see also the result [4, Proposition 5.10] concerning the linear setting). Thus, $\left\|\boldsymbol{M}^{-1/2}R(P)\right\|_{\mathrm{F}}$ is a good approximation of $\left\|\boldsymbol{M}^{1/2}\boldsymbol{\Pi}_P(P - P_*)\right\|_{\mathrm{F}}$, as shown on Figure 2 (central panel).

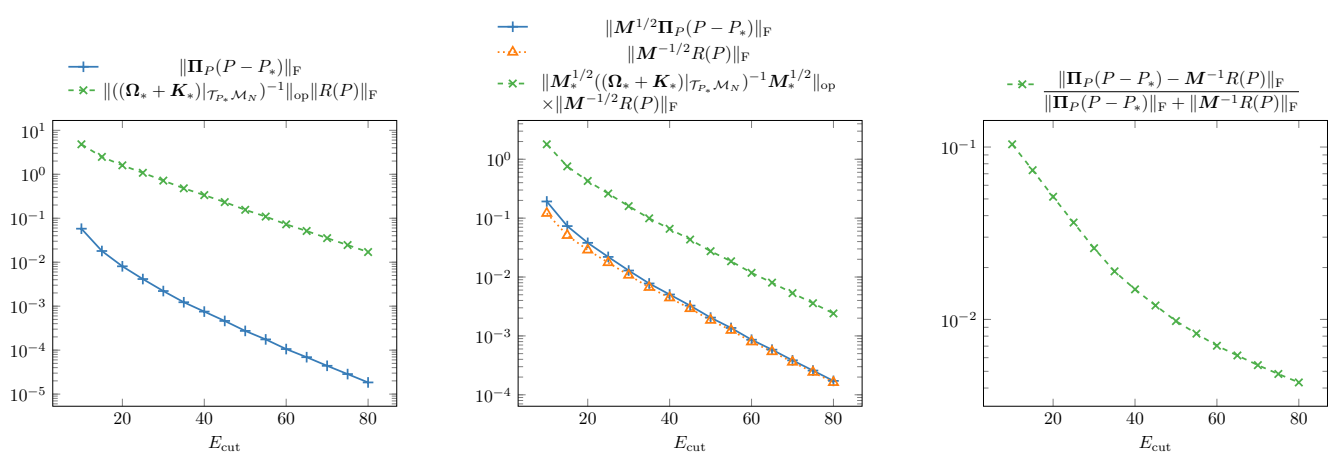

FIGURE 2 – Error bounds for Si based on (25) and (26). Left: L^2-norm; Center: H^1-type norm; Right: relative error between $\boldsymbol{\Pi}_P(P - P*)$ and $\boldsymbol{M}^{-1}R(P)$. It holds $\left\|((\boldsymbol{\Omega}_* + \boldsymbol{K}_*)|_{\mathcal{T}_{P_*}\mathcal{M}_\mathcal{N}})^{-1}\right\|_{\mathrm{op}} \approx 11.23$ and $\left\|\boldsymbol{M}_*^{1/2}((\boldsymbol{\Omega}_* + \boldsymbol{K}_*)|_{\mathcal{T}_{P_*}\mathcal{M}_\mathcal{N}})^{-1}\boldsymbol{M}_*^{1/2}\right\|_{\mathrm{op}} \approx 14.85$.

4.3. **Error bounds on QoIs and applications to interatomic forces.** Consider now a quantity of interest characterized by the smooth observable $A : \mathcal{M}_\mathcal{N} \to \mathcal{G}$, where $\mathcal{G}$ is a normed vector space (in particular, $\mathcal{G} = \mathbb{R}$ for

real QoIs such as the ground state energy, $\mathcal{G} = \mathbb{R}^{3N_{\rm at}}$ for interatomic forces, and, e.g., $\mathcal{G} = \mathrm{L}^2_\#$ for the ground-state densities). For such a QoI, there holds for $P \in \mathcal{M}_\mathcal{N}$ in the vicinity of P_*,

$$A(P) - A_* = \mathrm{d}A(P) \cdot (\mathbf{\Pi}_P(P - P_*)) + \text{h.o.t.}, \tag{27}$$

where $\mathrm{d}A(P) \in \mathcal{L}(\mathcal{T}_P\mathcal{M}_\mathcal{N}; \mathcal{G})$ is the derivative of A at P. We thus obtain the bound

$$\|A(P) - A_*\|_\mathcal{G} \leqslant \|\mathrm{d}A(P)\|_{\mathcal{T}_P\mathcal{M}_\mathcal{N} \to \mathcal{G}} \|\mathbf{\Pi}_P(P - P_*)\|_{\mathcal{T}_P\mathcal{M}_\mathcal{N}} \quad (+ \text{ h.o.t.}) \tag{28}$$

for given norms $\|\cdot\|_\mathcal{G}$ and $\|\cdot\|_{\mathcal{T}_P\mathcal{M}_\mathcal{N}}$ on $\mathcal{G}$ and $\mathcal{T}_P\mathcal{M}_\mathcal{N}$ respectively, and associated operator norm $\|\cdot\|_{\mathcal{T}_P\mathcal{M}_\mathcal{N} \to \mathcal{G}}$ on $\mathcal{L}(\mathcal{T}_P\mathcal{M}_\mathcal{N}; \mathcal{G})$.

Let us start with the simple case of the component of the force on atom j along the direction α due to the local part of the pseudopotential. Since this QoI is scalar, we have $\mathcal{G} = \mathbb{R}$. Using (20), we get

$$F^{\rm loc}_{j,\alpha}(P) = -\mathrm{Tr}\left(\frac{\partial V_{\rm loc}}{\partial X_{j,\alpha}} P\right).$$

Thus (28) becomes, using the Frobenius norm on $\mathcal{T}_P\mathcal{M}_\mathcal{N}$,

$$|F^{\rm loc}_{j,\alpha}(P) - F^{\rm loc}_{j,\alpha}(P_*)| \leqslant \left\|\mathbf{\Pi}_P \frac{\partial V_{\rm loc}}{\partial X_{j,\alpha}}\right\|_{\rm F} \|\mathbf{\Pi}_P(P - P_*)\|_{\rm F} \quad (+ \text{ h.o.t.}), \tag{29}$$

with

$$\mathbf{\Pi}_P \frac{\partial V_{\rm loc}}{\partial X_{j,\alpha}} \simeq_\Phi (1 - \Phi\Phi^*)\left(\frac{\partial V_{\rm loc}}{\partial X_{j,\alpha}} \Phi\right). \tag{30}$$

We plot in Figure 3 (left panel) the bound (29). The latter is pessimistic by more than three orders of magnitude, and its relative accuracy gets worse and worse as the cut-off energy increases.

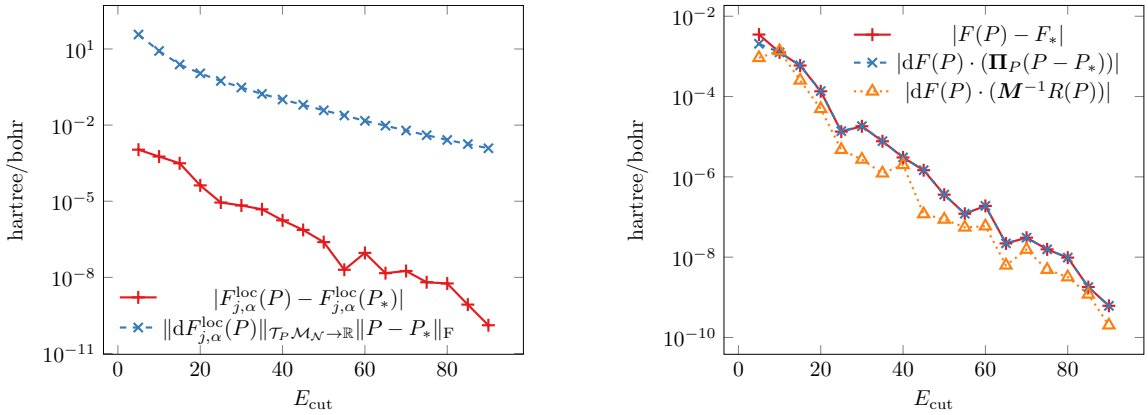

FIGURE 3 – Silicon. (Left panel) Inaccurate error bound (29) for the component of the force on atom $j = 1$ along direction $\alpha = (1, 0, 0)$ due to the local part of the pseudopotential. (Right panel) Approximation of $|F(P) - F_*|$ obtained by dropping the h.o.t. in the generic formula (27) and applying the derivative $\mathrm{d}F(P)$ either to the actual error $\mathbf{\Pi}_P(P - P_*)$ or the preconditioned residual $\boldsymbol{M}^{-1}R(P)$. The approximation $\mathrm{d}F(P) \cdot (\mathbf{\Pi}_P(P - P_*))$ matches asymptotically the error $F(P) - F_*$, validating again the rapid establishment of the linear regime. On the other hand, the approximation $\mathrm{d}F(P) \cdot (\boldsymbol{M}^{-1}R(P))$ does not match asymptotically.

The bound (29) using the operator norm of $\mathrm{d}A(P)$ being very inaccurate, we tested another approach consisting in using directly (27) to evaluate the error on the QoI by applying the derivative $\mathrm{d}A(P)$ to a computable approximation of $\mathbf{\Pi}_P(P - P_*)$. Relying on the results in the previous section showing that $\boldsymbol{M}^{-1/2}R(P)$ is a good approximation of $\boldsymbol{M}^{1/2}\mathbf{\Pi}_P(P - P_*)$ in Frobenius norm, it is tempting to replace $\mathbf{\Pi}_P(P - P_*)$ by $\boldsymbol{M}^{-1}R(P)$ in (27) and approximate $F(P) - F_*$ by $\mathrm{d}F(P) \cdot (\boldsymbol{M}^{-1}R(P))$, this approximation being justified by Figure 2 (right panel). Indeed the continuous counterpart of the asymptotic equivalence between $\boldsymbol{M}^{-1/2}R(P)$ and $\boldsymbol{M}^{1/2}\mathbf{\Pi}_P(P - P_*)$ for the Frobenius (L^2-type) norm is that the preconditioned residual and the error on the density matrix are asymptotically equivalent in H^1-type norms, while the derivative of the interatomic forces observable is continuous on H^1-type spaces. This idea is tested in Figure 3 (right panel). However, this leads to an underestimation of the error, although by a small factor. The reason is that even if $P - P_*$ and $M^{-1}R(P)$ do match asymptotically for the suitable norms, this is not the case for $\mathrm{d}F(P) \cdot (\mathbf{\Pi}_P(P - P_*))$ and $\mathrm{d}F(P) \cdot (\boldsymbol{M}^{-1}R(P))$ for reasons made clear in the next section.

Remark 3. In our simulations, the computation of $\mathrm{d}A(P) \cdot X$ for $X \in \mathcal{T}_P\mathcal{M}_\mathcal{N}$ is performed by forward-mode automatic differentiation using the `ForwardDiff.jl` Julia package [30].

We summarize the results of this section in Figure 4, displaying the combination of these bounds: the successive operator norms result in very inaccurate bounds (from six to eleven orders of magnitude) for the error on the forces.

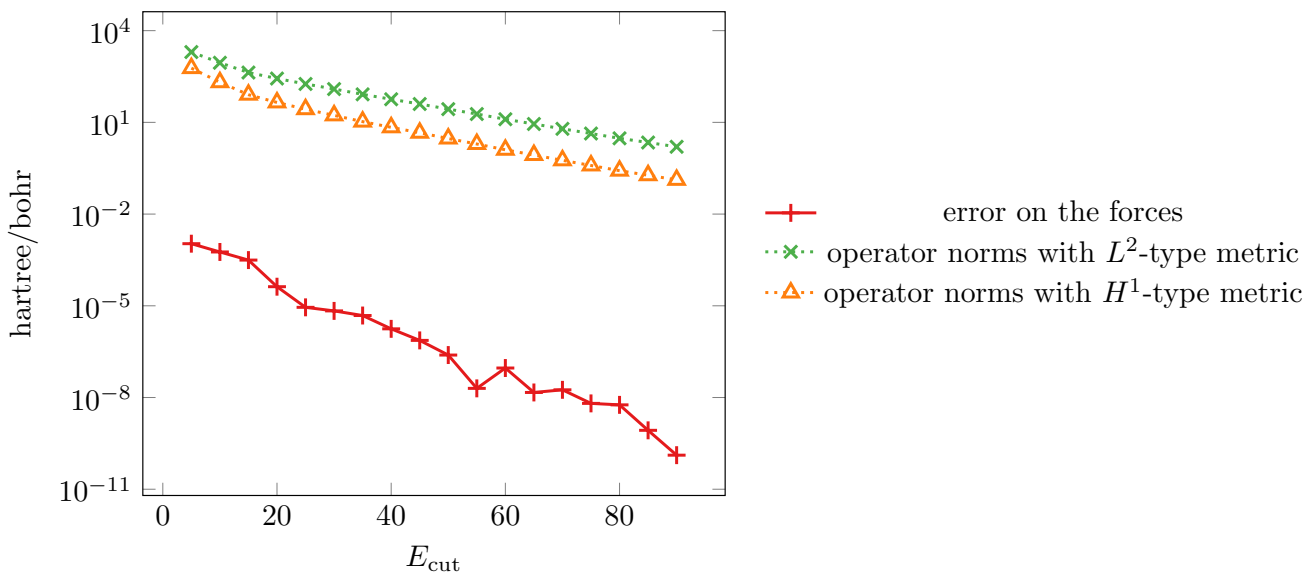

FIGURE 4 – Combination of the error estimate (28) on the interatomic forces with the error estimate on the error in L^2-type norm (25) and H^1-type norm (26). The inaccuracy of the bounds accumulates and results in extremely inaccurate bounds, from six to eleven orders of magnitude.

5. Improved error bounds based on frequencies splitting

5.1. **Spectral decomposition of the error.** In the previous section, we saw that even if $\mathbf{\Pi}_P(P - P_*)$ and $\boldsymbol{M}^{-1}R(P)$ are asymptotically equivalent in suitable norms, replacing the former by the latter in (27) when $A = F$ (interatomic forces) results in a large error, even in the asymptotic regime.

To analyze this issue, we use the decomposition

$$\mathcal{X}_{E_{\mathrm{cut,ref}}} = \mathcal{X}_{E_{\mathrm{cut}}} \oplus \mathcal{X}_{E_{\mathrm{cut}}}^{\perp}. \tag{31}$$

Since $\mathcal{X}_{E_{\mathrm{cut}}} = \mathrm{Span}\left(e_G,\ \frac{|G|^2}{2} \leqslant E_{\mathrm{cut}}\right)$ and $\mathcal{X}_{E_{\mathrm{cut}}}^{\perp} = \mathrm{Span}\left(e_G,\ E_{\mathrm{cut}} < \frac{|G|^2}{2} \leqslant E_{\mathrm{cut,ref}}\right)$, (31) corresponds to a low *vs* high frequency splitting. Using the identification of $\mathcal{X}_{E_{\mathrm{cut,ref}}} \equiv \mathbb{C}^\mathcal{N}$ introduced in subsection 3.2, (31) boils down to decomposing $\mathbb{C}^\mathcal{N}$ as

$$\mathbb{C}^\mathcal{N} = \mathcal{X} \oplus \mathcal{X}^\perp \quad \text{with} \quad \mathcal{X} = \begin{pmatrix} \mathbb{C}^{N_b} \\ 0_{\mathbb{C}^{\mathcal{N}-N_b}} \end{pmatrix} \quad \text{and} \quad \mathcal{X}^\perp = \begin{pmatrix} 0_{\mathbb{C}^{N_b}} \\ \mathbb{C}^{\mathcal{N}-N_b} \end{pmatrix}.$$

Let $\Phi \in \mathbb{C}^{\mathcal{N}\times N_{\mathrm{el}}}$ be such that $\Phi^*\Phi = I_{N_{\mathrm{el}}}$ and $P = \Phi\Phi^* \in \mathcal{M}_\mathcal{N}$. Combining the identification $\mathcal{X}_{E_{\mathrm{cut,ref}}} \equiv \mathbb{C}^\mathcal{N}$ described above with the relation (11) identifying a matrix X of the tangent space $\mathcal{T}_P\mathcal{M}_\mathcal{N}$ with a collection $\Xi = (\xi_1|\cdots|\xi_{N_{\mathrm{el}}}) \in \mathbb{C}^{\mathcal{N}\times N_{\mathrm{el}}}$ of orbital variations such that $\Phi^*\Xi = 0$, the decomposition (31) induces a decomposition of the tangent space $\mathcal{T}_P\mathcal{M}_\mathcal{N}$ into two orthogonal subspaces $\mathbf{\Pi}_{E_{\mathrm{cut}}}\mathcal{T}_P\mathcal{M}_\mathcal{N}$ and $\mathbf{\Pi}_{E_{\mathrm{cut}}}^{\perp}\mathcal{T}_P\mathcal{M}_\mathcal{N}$ (for the Frobenius inner product):

$$\mathbf{\Pi}_{E_{\mathrm{cut}}}\left(\sum_{i=1}^{N_{\mathrm{el}}} |\phi_i\rangle\langle\xi_i| + |\xi_i\rangle\langle\phi_i|\right) := \sum_{i=1}^{N_{\mathrm{el}}} |\phi_i\rangle\langle\Pi_\mathcal{X}\xi_i| + |\Pi_\mathcal{X}\xi_i\rangle\langle\phi_i|,$$

$$\mathbf{\Pi}_{E_{\mathrm{cut}}}^{\perp}\left(\sum_{i=1}^{N_{\mathrm{el}}} |\phi_i\rangle\langle\xi_i| + |\xi_i\rangle\langle\phi_i|\right) := \sum_{i=1}^{N_{\mathrm{el}}} |\phi_i\rangle\langle\Pi_\mathcal{X}^{\perp}\xi_i| + |\Pi_\mathcal{X}^{\perp}\xi_i\rangle\langle\phi_i|,$$

where $\Pi_{\mathcal{X}}$ is the orthogonal projector on $\mathcal{X}$ (for the canonical inner product of $\mathbb{C}^{\mathcal{N}}$) and $\Pi_{\mathcal{X}}^{\perp} = 1 - \Pi_{\mathcal{X}}$. If P solves the minimization problem (4), we infer from the first-order optimality conditions that the residual $R(P)$ is orthogonal to $\mathbf{\Pi}_{E_{\rm cut}}\mathcal{T}_P\mathcal{M}_{\mathcal{N}}$, meaning that the vectors $r_i(P)$ such that

$$R(P) = \sum_{i=1}^{N_{\rm el}} |\phi_i\rangle \langle r_i(P)| + |r_i(P)\rangle \langle \phi_i|$$

belong to $\mathcal{X}^{\perp}$. Note that in practice, this is not exactly true for the full Kohn–Sham model because of the numerical quadrature errors involved in the treatment of the exchange-correlation terms.

Now $P - P_* \approx ((\mathbf{\Omega}(P) + \boldsymbol{K}(P))|_{\mathcal{T}_P\mathcal{M}_{\mathcal{N}}})^{-1}R(P)$ contains two components: one in $\mathbf{\Pi}_{E_{\rm cut}}\mathcal{T}_P\mathcal{M}_{\mathcal{N}}$ and one in $\mathbf{\Pi}_{E_{\rm cut}}^{\perp}\mathcal{T}_P\mathcal{M}_{\mathcal{N}}$. In the high-frequency subspace $\mathbf{\Pi}_{E_{\rm cut}}^{\perp}\mathcal{T}_P\mathcal{M}_{\mathcal{N}}$, the leading term in $(\mathbf{\Omega}(P) + \boldsymbol{K}(P))|_{\mathcal{T}_P\mathcal{M}_{\mathcal{N}}}$ comes from the contribution of the Laplacian arising in the Hamiltonian h_0, which is well approximated by the super-operator $\boldsymbol{M}$. This claim is supported by Proposition 1, in which we prove in a simplified setting that $((\mathbf{\Omega}(P) + \boldsymbol{K}(P))|_{\mathcal{T}_P\mathcal{M}_{\mathcal{N}}})^{-1}\mathbf{\Pi}_{E_{\rm cut}}^{\perp}$ is asymptotically equivalent to $\boldsymbol{M}^{-1}\mathbf{\Pi}_{E_{\rm cut}}^{\perp}$.

This is what we observe in Figure 5 (central and right panels): if P is the solution to (4), the residual $R(P)$ is supported in $\mathbf{\Pi}_{E_{\rm cut}}^{\perp}\mathcal{T}_P\mathcal{M}_{\mathcal{N}}$ (up to numerical quadrature errors). In accordance with Proposition 1, the difference between the error $P - P_* \approx \mathbf{\Pi}_P(P - P_*)$ and the preconditioned residual $\boldsymbol{M}^{-1}R(P)$ (Figure 6) is smaller in Frobenius norm than the preconditioned residual itself. This explains our observations in subsection 4.2 that $\|P - P_*\|_{\rm F}$ is well approximated by $\left\|\boldsymbol{M}^{-1}R(P)\right\|_{\rm F}$. However, this does not imply that $\mathrm{d}F(P)\cdot\mathbf{\Pi}_P(P - P_*)$ is well-approximated by $\mathrm{d}F(P)\cdot(\boldsymbol{M}^{-1}R(P))$. This is because the gradients $\nabla F_{j,\alpha}(P)$ are mostly supported on low frequencies, as illustrated in Figure 5 (left panel). Although the low-frequency contribution to the error $\mathbf{\Pi}_P(P - P_*)$ is of smaller magnitude than the high-frequency contribution, its contribution to $\mathrm{d}F(P)\cdot(\mathbf{\Pi}_P(P - P_*))$ is very significant. The fact that the low-frequency error is not captured at all by the purely high-frequency term $\boldsymbol{M}^{-1}R(P)$ is responsible for the poor approximation of the error $F(P) - F_*$ by $\mathrm{d}F(P)\cdot(\boldsymbol{M}^{-1}R(P))$.

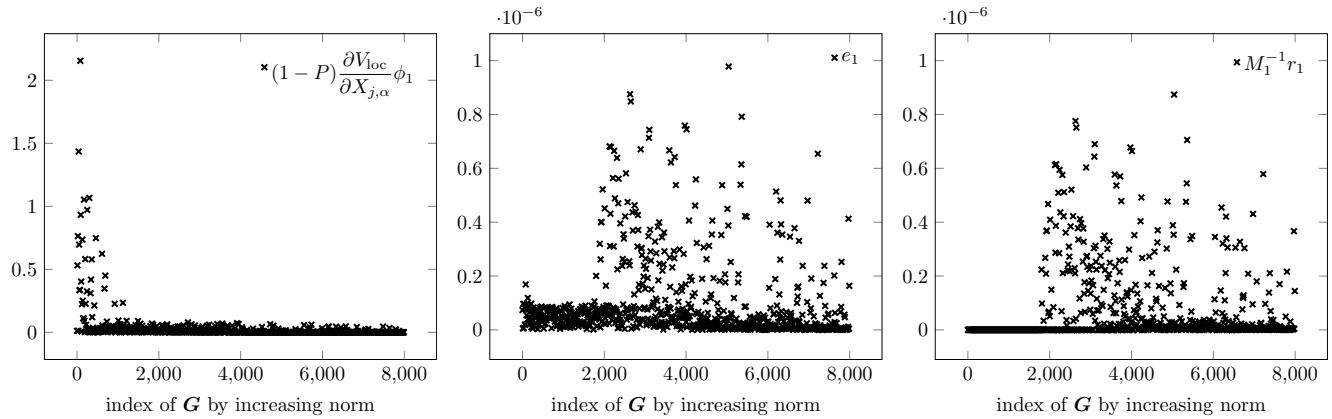

FIGURE 5 – Fourier coefficients moduli in the orbital representation $\mathbf{\Pi}_P(P - P_*) \simeq_{\Phi} (e_i)_{1\leqslant i\leqslant N}$ and $\boldsymbol{M}^{-1}R(P) \simeq_{\Phi} (M_i^{-1}r_i)_{1\leqslant i\leqslant N}$. (Left) Test function $(1-P)\frac{\partial V_{\rm loc}}{\partial X_{j,\alpha}}\phi_1$ (see (30)).

Now that we have understood the reason why it is not possible to approximate the error $F(P) - F_*$ on the interatomic forces by the computable term $\mathrm{d}F(P)\cdot(\boldsymbol{M}^{-1}R(P))$, we propose in the next section a way to evaluate this error, based on the linearization (23) and the frequencies splitting we just introduced.

5.2. **Improving the error estimation.** We now decompose tangent vectors and operators according to the splitting $\mathbf{\Pi}_{E_{\rm cut}}\mathcal{T}_P\mathcal{M}_{\mathcal{N}}$ and $\mathbf{\Pi}_{E_{\rm cut}}^{\perp}\mathcal{T}_P\mathcal{M}_{\mathcal{N}}$, which we respectively label by 1 and 2 for simplicity. In this way, the error-residual relationship can be written in concise form with obvious notation as

$$\begin{bmatrix} (\mathbf{\Omega}+\boldsymbol{K})_{11} & (\mathbf{\Omega}+\boldsymbol{K})_{12} \\ (\mathbf{\Omega}+\boldsymbol{K})_{21} & (\mathbf{\Omega}+\boldsymbol{K})_{22} \end{bmatrix} \begin{bmatrix} P_1 - P_{*1} \\ P_2 - P_{*2} \end{bmatrix} = \begin{bmatrix} R_1 \\ R_2 \end{bmatrix}.$$

Recall that $(\mathbf{\Omega}(P) + \boldsymbol{K}(P))|_{\mathcal{T}_P\mathcal{M}_{\mathcal{N}}}$ is only invertible at high cost as it has a priori nonzero values on the four components of the operator arising from the low frequencies/high frequencies splitting of the operator. The computational cost for the inversion is equivalent to performing a Newton step on the reference grid. But we can

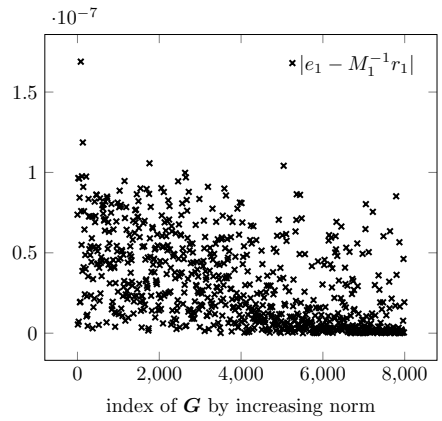

FIGURE 6 – Fourier coefficients of the difference between the error e_1 and the preconditioned residual $M_1^{-1} r_1$, where $\boldsymbol{\Pi}_P(P - P_*) \simeq_\Phi (e_i)_{1 \leqslant i \leqslant N}$ and $\boldsymbol{M}^{-1} R(P) \simeq_\Phi (M_i^{-1} r_i)_{1 \leqslant i \leqslant N}$. Low frequencies contribute greatly.

make approximations to invert it only on the coarse grid $\mathcal{X}_{E_{\mathrm{cut}}}$ and approximate the low frequency error components. In the same spirit as for the perturbation-theory based post-processing method introduced in [5, 9] and the Feshbach-Schur method analyzed in [11], we make the following approximations:

$$(\boldsymbol{\Omega} + \boldsymbol{K})_{21} \approx 0 \quad \text{and} \quad (\boldsymbol{\Omega} + \boldsymbol{K})_{22} \approx \boldsymbol{M}_{22},$$

which yields

$$\begin{bmatrix} (\boldsymbol{\Omega} + \boldsymbol{K})_{11} & (\boldsymbol{\Omega} + \boldsymbol{K})_{12} \\ 0 & \boldsymbol{M}_{22} \end{bmatrix} \begin{bmatrix} P_1 - P_{*1} \\ P_2 - P_{*2} \end{bmatrix} = \begin{bmatrix} R_1 \\ R_2 \end{bmatrix}$$

and therefore

$$P_2 - P_{*2} \approx \boldsymbol{M}_{22}^{-1} R_2, \tag{32}$$

$$P_1 - P_{*1} \approx (\boldsymbol{\Omega} + \boldsymbol{K})_{11}^{-1} \left(R_1 - (\boldsymbol{\Omega} + \boldsymbol{K})_{12} \boldsymbol{M}_{22}^{-1} R_2\right). \tag{33}$$

This requires only a single inexpensive computation on the fine grid. The main bottleneck is then to solve a linear system with operator $(\boldsymbol{\Omega} + \boldsymbol{K})_{11}$, which is as expensive as a full Newton step on the coarse grid $\mathcal{X}_{E_{\mathrm{cut}}}$. Since $R_1 = 0$ when P is the optimal Galerkin solution on $\mathcal{X}_{E_{\mathrm{cut}}}$, we can understand the previous attempt to replace $P - P_*$ by $\boldsymbol{M}^{-1} R(P)$ as (32). Not neglecting $(\boldsymbol{\Omega} + \boldsymbol{K})_{12}$ in (33) gives rise to a correction on the coarse space also. We denote by $R_{\mathrm{Schur}}(P)$ the new residual

$$R_{\mathrm{Schur}}(P) = \begin{bmatrix} (\boldsymbol{\Omega} + \boldsymbol{K})_{11}^{-1} \left(R_1 - (\boldsymbol{\Omega} + \boldsymbol{K})_{12} \boldsymbol{M}_{22}^{-1} R_2\right) \\ \boldsymbol{M}_{22}^{-1} R_2 \end{bmatrix}.$$

To illustrate the validity of these approximations, we plotted in Figure 7 the components of r_{Schur}, the orbital representation of R_{Schur}. We see that this time, the error is well approximated by (33) in the low-frequency space.

In Figure 8, we plot the new estimate $\mathrm{d}F(P) \cdot (R_{\mathrm{Schur}}(P))$ of the error $F(P) - F_*$ as well as the differences

$$F_{\mathrm{err}} - F_* \coloneqq F(P) - \mathrm{d}F(P) \cdot (\boldsymbol{\Pi}_P(P - P_*)) - F_*,$$
$$F_{\mathrm{res}} - F_* \coloneqq F(P) - \mathrm{d}F(P) \cdot (\boldsymbol{M}^{-1} R(P)) - F_*,$$
$$F_{\mathrm{Schur}} - F_* \coloneqq F(P) - \mathrm{d}F(P) \cdot (R_{\mathrm{Schur}}(P)) - F_*,$$

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

Remark 4. The quantity $\mathrm{d}F(P) \cdot (R_{\mathrm{Schur}}(P))$ does not yield a guaranteed estimator of the error on the forces as it is obtained after several approximations and is only valid in the asymptotic regime. However, it can be computed for a cost comparable to the one of performing a SCF step on the same grid and can be used for two main purposes:

- as an error bound, as the error $F(P) - F_*$ is reasonably well approximated by $\mathrm{d}F(P) \cdot (R_{\mathrm{Schur}}(P))$;

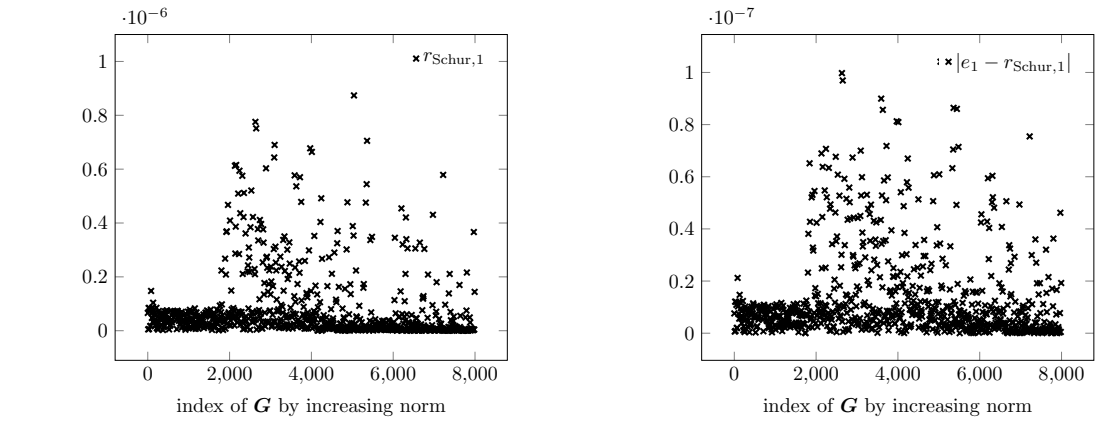

FIGURE 7 – Fourier coefficients of the new residual $r_{\mathrm{Schur},1}$ and its comparison to the error e_1, where $\boldsymbol{\Pi}_P(P - P_*) \simeq_\Phi (e_i)_{1\leqslant i\leqslant N}$ and $R_{\mathrm{Schur}}(P) \simeq_\Phi (r_{\mathrm{Schur},i})_{1\leqslant i\leqslant N}$. (Left) Components of the modified residual $r_{\mathrm{Schur},1}$. (Right) Difference between the error and the new residual: low frequencies are better approximated (compare with Figure 6).

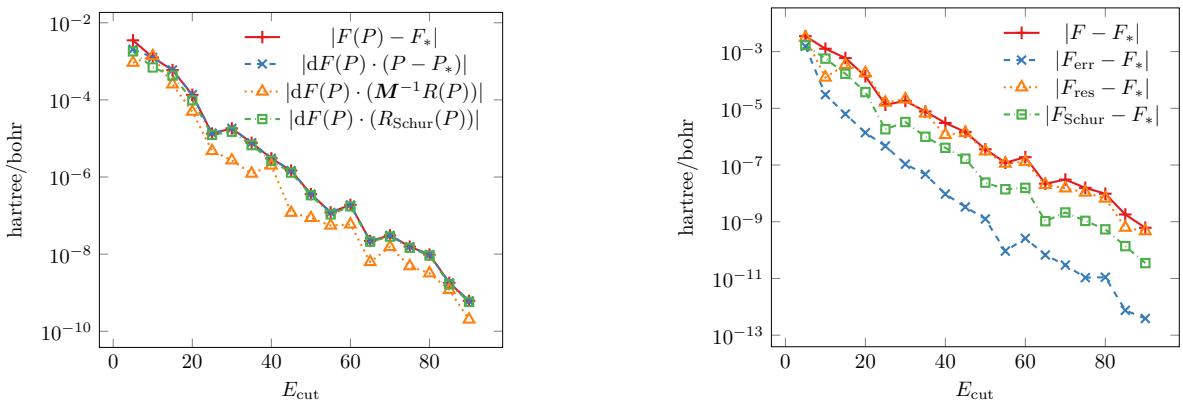

FIGURE 8 – (Left) Estimation of the error $F(P)-F_*$ with $\mathrm{d}F(P)\cdot X$ where X is either the exact error $\boldsymbol{\Pi}_P(P-P_*)$, the preconditioned residual $\boldsymbol{M}^{-1}R(P)$ or the modified residual $R_{\mathrm{Schur}}(P)$. (Right) Enhancement of the estimation of the forces by replacing $F(P)$ with $F(P) - \mathrm{d}F(P)\cdot X$ where X is either the exact error $\boldsymbol{\Pi}_P(P - P_*)$, the preconditioned residual $\boldsymbol{M}^{-1}R(P)$ or the modified residual $R_{\mathrm{Schur}}(P)$.

- as a more precise approximation of the QoI, as the forces $F_j(P)$ on atom j obtained by a variational approximation on a coarse grid are improved by the post-processing $F_j(P) \mapsto F_j(P) - \mathrm{d}F_j(P)\cdot(R_{\mathrm{Schur}}(P))$.

6. Numerical examples with more complex systems

We perform the same simulations as for silicon, but for more complex systems, namely GaAs and TiO_2. The calculations are still performed within the LDA approximation with GTH pseudopotentials and Teter 93 exchange-correlation functional, with a $2\times 2\times 2$ k-point grid to discretize the Brillouin zone, and the reference solutions are obtained for $E_{\mathrm{cut,ref}} = 125$ Ha. We describe here the numerical setting for both systems.

GaAs: We use the usual periodic lattice for the FCC phase of GaAs, with lattice constant 10.68 bohrs, close to but not exactly at the equilibrium configuration in order to get nonzero forces. The Ga atom is placed at fractional coordinates $(\frac{1}{8}, \frac{1}{8}, \frac{1}{8})$ and the As atom at fractional coordinates $(-\frac{1}{8}, -\frac{1}{8}, -\frac{1}{8})$. The Ga atom is then displaced by $\frac{1}{15}(0.24, -0.33, 0.12)$ to get nonzero forces. In this setting, the reference values for the energy is $E_* = -8.572$ Ha and the interatomic forces are, in hartree/bohr,

$$F_* = \begin{bmatrix} -0.0448 & 0.0448 \\ 0.0722 & -0.0722 \\ -0.0251 & 0.0251 \end{bmatrix},$$

where the first column are the forces acting on the Ga atom in each direction, and the second column are the forces acting on the As atom.

TiO_2: We use the MP-2657 configuration in the primitive cell from the Materials Project [28]. We apply the small displacement $\frac{1}{5}(0.22, -0.28, 0.35)$ to the equilibrium position of the first Ti atom to get nonzero forces. In this setting, the reference values for the energy is $E_* = -71.589$ Ha and the interatomic forces are, in hartree/bohr,

$$F_* = \begin{bmatrix} -2.88 & 0.641 & 3.80 & 0.753 & -1.57 & -0.745 \\ 3.10 & -0.919 & -3.09 & -1.45 & 0.800 & 1.56 \\ 0.136 & 0.403 & -0.368 & -0.786 & 0.251 & 0.364 \end{bmatrix},$$

where the first two columns are the forces acting on the two Ti atoms in each direction, and the other columns are the forces acting on the four O atoms.

We plot in Figure 9 the energy, density and forces obtained after a Newton step on the fine grid starting from the variational solution on the coarse grid given by E_{cut}, for GaAs and TiO_2. The fast establishment of the asymptotic regime is confirmed for the two new systems as, even for small E_{cut}'s, the so-obtained QoIs are orders of magnitude more accurate than the ones obtained by the variational solution on the coarse grid.

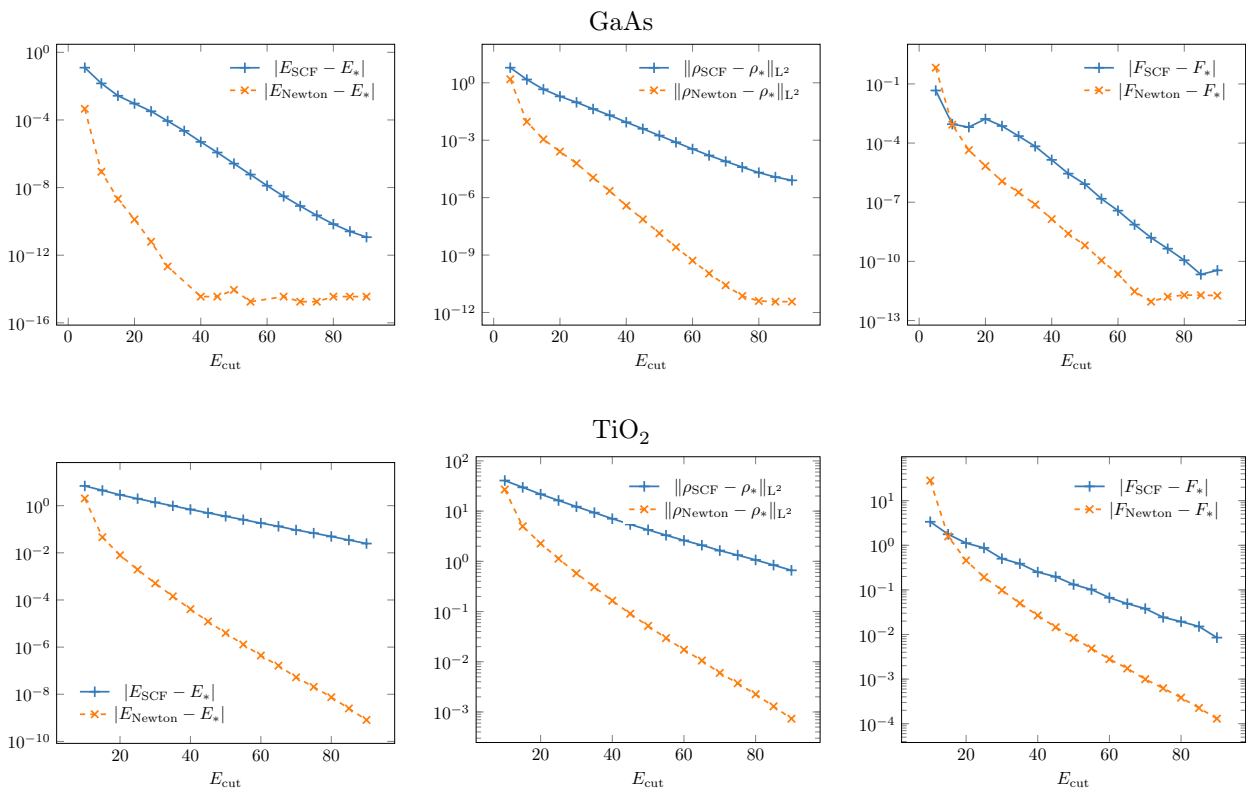

FIGURE 9 – Errors of some QoI as functions of E_{cut} (reference solution is obtained with $E_{\text{cut,ref}} = 125$ Ha) for GaAS and TiO_2. Solid lines: errors obtained with the variational solution in the space $\mathcal{X}_{E_{\text{cut}}}$. Dashed lines: errors obtained with one Newton step on the reference grid, starting from the variational solution in the space $\mathcal{X}_{E_{\text{cut}}}$. Left panel: energy (hartree), central panel: discrete L^2 norm of the density, right panel: interatomic forces (hartree/bohr). To be compared with Figure 1.

We plot in Figure 10 the estimation of the actual error $F(P) - F_*$ with $\mathrm{d}F(P) \cdot X$ where X is either $\mathbf{\Pi}_P(P - P_*)$, $R(P)$ or $R_{\text{Schur}}(P)$. In Figure 11, we plot the improvement of the estimation of the forces $F(P) - \mathrm{d}F(P) \cdot X$ where X is either $\mathbf{\Pi}_P(P - P_*)$, $R(P)$ or $R_{\text{Schur}}(P)$. Just as for silicon, the estimation is well improved with the modified residual R_{Schur}. Note that in the GaAs case, there is a plateau for high E_{cut}'s. This phenomenon is explained in the remark below.

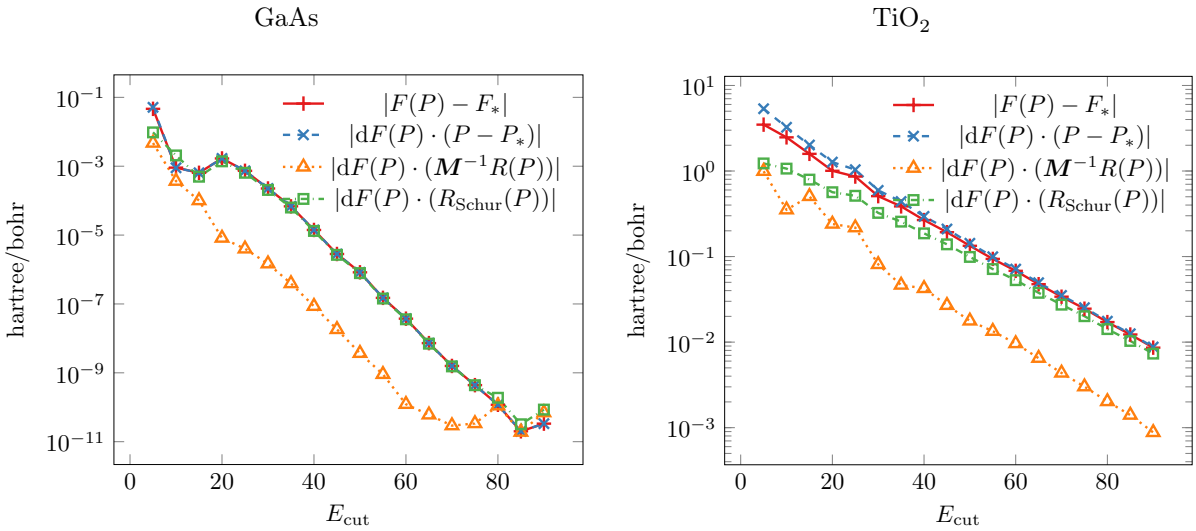

FIGURE 10 – Estimation of the error $F(P)-F_*$ with $\mathrm{d}F(P)\cdot X$ where X is either the exact error $\mathbf{\Pi}_P(P-P_*)$, the preconditioned residual $\boldsymbol{M}^{-1}R(P)$ or the modified residual $R_{\text{Schur}}(P)$. To be compared with Figure 8 (Left).

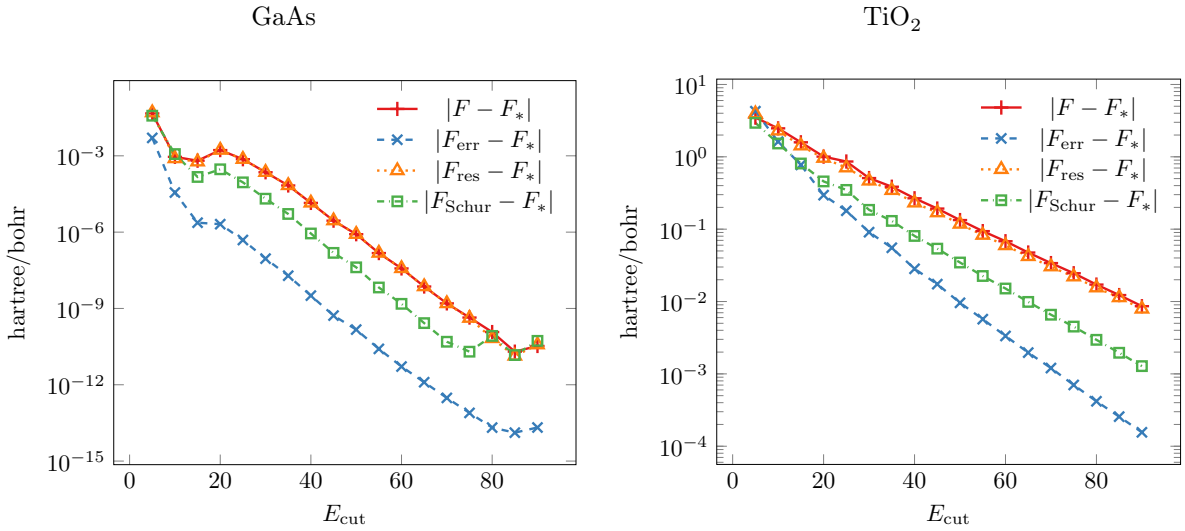

FIGURE 11 – Enhancement of the estimation of the forces by replacing $F(P)$ with $F(P)-\mathrm{d}F(P)\cdot X$ where X is either the exact error $\mathbf{\Pi}_P(P-P_*)$, the preconditioned residual $\boldsymbol{M}^{-1}R(P)$ or the modified residual $R_{\text{Schur}}(P)$. To be compared with Figure 8 (Right).

Remark 5. The plateau observed Figure 10 and Figure 11 for GaAs comes from the numerical quadrature scheme used to compute the exchange-correlation energy and the corresponding matrix elements. In fact, we also observed such plateaus for silicon and TiO_2 with the default quadrature scheme of DFTK, but these disappeared by using 8 times as many numerical quadrature points. With this more accurate numerical quadrature scheme, the plateau for GaAs is lower but still visible. It disappears when further increasing the number of quadrature points, at the price of longer computations.

7. Conclusion

In this work, we have investigated methods to estimate the error on interatomic forces resulting from plane-wave discretizations of the Kohn–Sham equations. On the systems we investigated, we find the following:

- Linearizing the equations around a solution is a good approximation, even for energy cut-offs as small as 5 hartree (Figure 1).
- The naive approach based on the computation of operator norms proves to be extremely inefficient, overestimating the error by several orders of magnitude. This is the case even when using appropriate H^1-type norms (Figure 4). The reason is that the discretization error is mostly made up of high frequency components, whereas quantities of interest are mostly supported on low frequencies, resulting in very suboptimal Cauchy-Schwarz inequalities (Figure 5).
- Replacing directly the error by the preconditioned residual yields reasonable estimates of the errors, but they are not systematic upper bounds (Figure 3).
- A Schur approach based on a low/high frequency splitting systematically improves the solution and gives reliable estimates of the error (Figure 8), at the price of more computational work.

Our results validate on realistic test cases and for properties such as interatomic forces the frequency splitting approach already introduced in [5, 9, 11].Thanks to the modular nature of DFTK and the use of automatic differentiation, the implementation of our estimates is relatively simple and convenient. It is publicly available at `https://github.com/gkemlin/paper-forces-estimator`. The algorithm proceeds in two steps: i) the computation of the residual on the fine grid and ii) a linear system solve involving the Jacobian on the coarse grid. The computational cost of step i) is negligible compared to that of step ii), which is roughly that of a full self-consistent computation on the coarse grid. Therefore, for roughly twice the cost of a standard computation, one obtains an accurate approximation of the discretization error on the interatomic forces (or, equivalently, a better estimate of the latter).

The scope of this work is limited to gapped systems at zero temperature and to the study of the discretization error. Interesting perspectives for future work include the application of this methodology to the error resulting from an incomplete self-consistent cycle, and to finite-temperature models, including metals (see [16] for an extension of the linearized equations to the finite-temperature case).

Acknowledgements

The authors would like to thank Michael F. Herbst for fruitful discussions and help with DFTK, and Niklas Schmitz and Markus Towara for the implementation of forward differentiation into DFTK. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 810367). G.D. acknowledges the support of the Ecole des Ponts-ParisTech, as this work was carried out in the framework of an associated researcher position at CERMICS.

Appendix A. Mathematical justification

The purpose of this appendix is to explain mathematically in a simplified setting the observation in subsection 4.2 that $\left\|\boldsymbol{M}^{-1/2}\boldsymbol{\Pi}_P R(P)\right\|_{\mathrm{F}}$ was a good approximation of $\left\|\boldsymbol{M}^{1/2}\boldsymbol{\Pi}_P(P-P_*)\right\|_{\mathrm{F}}$. For this purpose, we work in a slightly different framework than the one we used in the rest of the paper, and consider the infinite-dimensional version of Problem (3) associated with the periodic Gross–Pitaevskii model in dimension $d \leqslant 3$, which reads as

$$E_* \coloneqq \min\{E(P),\ P \in \mathcal{M}_\infty\}, \tag{34}$$

with $\mathcal{M}_\infty \coloneqq \{P \in \mathcal{S}(\mathrm{L}^2_\#) \,|\, P^2 = P,\ \mathrm{Tr}(P) = 1,\ \mathrm{Ran}\,(P) \subset \mathrm{H}^1_\#\}$ and $E(P) \coloneqq \mathrm{Tr}((-\Delta + V)P) + \frac{1}{2}\int_\Gamma \rho_P^2$. Here $\mathcal{S}(\mathrm{L}^2_\#)$ denotes the space of self-adjoint operators on $\mathrm{L}^2_\#$, V a given function of $\mathrm{L}^\infty_\#$, and ρ_P the density of P. The condition $\mathrm{Ran}\,(P) \subset \mathrm{H}^1_\#$ ensures that both the linear and nonlinear terms in the energy functional $E(P)$ are well-defined and finite. It is convenient to rewrite (34) in the orbital framework. Any state $P \in \mathcal{M}_\infty$ is rank-1 and such that $\mathrm{Ran}\,(P) \subset \mathrm{H}^1_\#$. It can therefore be represented by a function $\phi \in \mathrm{H}^1_\#$ such that $\|\phi\|_{\mathrm{L}^2_\#} = 1$ through the relation $P = |\phi\rangle\langle\phi|$ (using Dirac's notation). The orbital formulation of problem (34) reads

$$E_* \coloneqq \min\{\mathcal{E}^{\mathrm{GP}}(\phi),\ \phi \in \mathrm{H}^1_\#,\ \|\phi\|_{\mathrm{L}^2_\#} = 1\}, \tag{35}$$

with $\mathcal{E}^{\mathrm{GP}}(\phi) \coloneqq \int_\Gamma |\nabla\phi|^2 + \int_\Gamma V|\phi|^2 + \frac{1}{2}\int_\Gamma |\phi|^4$. It is well-known (see e.g. the Appendix of [1]) that the minimizer of (34) is unique, and that the set of solutions of (35) is $(e^{i\alpha}\phi_*)_{\alpha\in\mathbb{R}}$, where $(\lambda_*, \phi_*) \in \mathbb{R} \times \mathrm{H}^1_\#$ is the unique solution to

$$\begin{cases} -\Delta\phi_* + V\phi_* + \phi_*^3 = \lambda_*\phi_*, \\ \|\phi_*\|_{\mathrm{L}^2_\#} = 1, \quad \phi_* > 0 \text{ on } \mathbb{R}^d. \end{cases} \tag{36}$$

We consider the variational approximation of (34) in the finite dimensional space

$$\mathcal{X}_N := \mathrm{Span}\left(e_G,\ |G|^2/2 \leqslant N\right)$$

corresponding to a plane-wave discretization with energy cut-off $E_{\mathrm{cut}} = N$. We denote by Π_N the $\mathrm{L}^2_\#$-orthogonal projector on $\mathcal{X}_N$ and by $\Pi_N^\perp := 1 - \Pi_N$. For N large enough, the approximate ground-state P_N is unique and can be represented by a unique function ϕ_N real-valued and positive on $\mathbb{R}^3$ (see [1]), and it holds

$$\begin{cases} -\Delta\phi_N + \Pi_N\left(V\phi_N - \phi_N^3\right) = \lambda_N\phi_N, \\ \|\phi_N\|_{\mathrm{L}^2_\#} = 1, \end{cases}$$

for some uniquely defined $\lambda_N \in \mathbb{R}$. In addition, we have $\phi_* \in \mathrm{H}^2_\#$ and

$$\|\phi_N - \phi_*\|_{\mathrm{H}^2_\#} \underset{N\to\infty}{\longrightarrow} 0 \quad \text{and} \quad |\lambda_N - \lambda_*| \underset{N\to\infty}{\longrightarrow} 0. \tag{37}$$

Using similar notation as the one used in the rest of the paper, we introduce the following quantities:

- $\Pi_{\phi_N}^\perp$ is the orthogonal projector (for the $\mathrm{L}^2_\#$ inner product) onto $\phi_N^\perp$;
- A_N is the self-adjoint operator on $\phi_N^\perp$ defined by

$$A_N := (\Omega_N + K_N) \tag{38}$$

  where Ω_N and K_N represent, in the orbital framework, the super-operators $\boldsymbol{\Omega}(P_N)|_{T_{P_N}\mathcal{M}_\infty}$ and $\boldsymbol{K}(P_N)|_{T_{P_N}\mathcal{M}_\infty}$. We have

$$\forall\ \psi_N \in \phi_N^\perp, \quad \Omega_N\psi_N = \Pi_{\phi_N}^\perp\left(-\Delta + V + \phi_N^2 - \lambda_N\right)\psi_N, \tag{39}$$

$$\forall\ \psi_N \in \phi_N^\perp, \quad K_N\psi_N = \Pi_{\phi_N}^\perp\left(2\phi_N^2\psi_N\right); \tag{40}$$

- $M_N^{1/2}$ is the restriction of the operator $\Pi_{\phi_N}^\perp(1-\Delta)^{1/2}\Pi_{\phi_N}^\perp$ to the invariant subspace $\phi_N^\perp$.

We then have the following result, which justifies in this case the claim made in subsection 4.2 that $\boldsymbol{M}^{-1/2}\boldsymbol{M}^{1/2}((\boldsymbol{\Omega}(P)+\boldsymbol{K}(P))|_{\mathcal{T}_P\mathcal{M}_\mathcal{N}})^{-1}\boldsymbol{M}^{1/2}$ is close to identity on the subspace of high-frequency Fourier modes. It also justifies that $\boldsymbol{\Pi}_P(P-P_*) \approx \boldsymbol{M}^{-1}R(P)$ as $\boldsymbol{M}^{-1/2}$ is a uniformly bounded operator and $\boldsymbol{M}^{-1/2}R(P)$ is high-frequency:

$$\begin{aligned}\boldsymbol{\Pi}_P(P-P_*) &\approx \boldsymbol{M}^{-1/2}\boldsymbol{M}^{1/2}((\boldsymbol{\Omega}(P)+\boldsymbol{K}(P))|_{\mathcal{T}_P\mathcal{M}_\mathcal{N}})^{-1}\boldsymbol{M}^{1/2}\boldsymbol{M}^{-1/2}R(P) \\ &\approx \boldsymbol{M}^{-1/2}\boldsymbol{M}^{-1/2}R(P) = \boldsymbol{M}^{-1}R(P).\end{aligned}$$

Proposition 1. *We have*

$$\lim_{N\to\infty}\left\|M_N^{1/2}(\Omega_N + K_N)^{-1}M_N^{1/2} - I_{\mathcal{X}_N^\perp}\right\|_{\mathcal{X}_N^\perp\to\mathrm{L}^2_\#} = 0.$$

Proof. Let $W_N := V + 3\phi_N^2 - \lambda_N - 1$ and $W_* := V + 3\phi_*^2 - \lambda_* - 1$. In view of (37), W_N converges to W_* in $\mathrm{L}^\infty_\#$ when N goes to infinity. It also follows from [1, Lemma 1] that the self-adjoint operator

$$\begin{aligned}\widetilde{A}_* &:= -\Delta + V + 3\phi_*^2 - \lambda_* = (1-\Delta) + W_* \\ &= (1-\Delta)^{1/2}\Big(1 + (1-\Delta)^{-1/2}W_*(1-\Delta)^{-1/2}\Big)(1-\Delta)^{1/2}\end{aligned}$$

is coercive, hence, by the Lax–Milgram lemma, defines a continuous isomorphism from $\mathrm{H}^1_\#$ to $\mathrm{H}^{-1}_\#$. We denote by $\widetilde{A}_*^{-1}$ its inverse, seen as a bounded operator from $\mathrm{H}^{-1}_\#$ to $\mathrm{H}^1_\#$, so that $B_* := (1-\Delta)^{1/2}\widetilde{A}_*^{-1}(1-\Delta)^{1/2}$ defines a bounded operator on $\mathrm{L}^2_\#$.

Using the convergence results (37) and standard perturbation theory it follows that for N large enough, the operator $B_N := (1-\Delta)^{1/2}\widetilde{A}_N^{-1}(1-\Delta)^{1/2}$, where $\widetilde{A}_N := (1-\Delta) + W_N$, is bounded on $\mathrm{L}^2_\#$ uniformly in N, and that we have

$$B_N = \Big(1 + (1-\Delta)^{-1/2}W_N(1-\Delta)^{-1/2}\Big)^{-1} = 1 - B_N(1-\Delta)^{-1/2}W_N(1-\Delta)^{-1/2}. \tag{41}$$

We now compute the action of the operator $M_N^{1/2}A_N^{-1}M_N^{1/2} : \phi_N^\perp \to \phi_N^\perp$, relating it to $\widetilde{A}_N^{-1}$ and B_N, with A_N defined in (38). Let $\xi_N \in X_N^\perp$. As $\phi_N \in X_N$, we have $X_N^\perp \subset \phi_N^\perp$ so that $\xi_N \in X_N^\perp \subset \phi_N^\perp$, and $M_N^{1/2}\xi_N = (1-\Delta)^{1/2}\xi_N \in X_N^\perp \subset \phi_N^\perp$, where we used that X_N and $X_N^\perp$ are invariant subspaces of the operator $(1-\Delta)^{1/2}$. Let $v_N := A_N^{-1}M_N^{1/2}\xi_N = A_N^{-1}(1-\Delta)^{1/2}\xi_N \in \phi_N^\perp$. Using (39) and (40), we get

$$\Pi_{\phi_N}^\perp\left(-\Delta + V + 3\phi_N^2 - \lambda_N\right)v_N = (1-\Delta)^{1/2}\xi_N, \quad \textit{i.e.} \quad \Pi_{\phi_N}^\perp\widetilde{A}_N v_N = (1-\Delta)^{1/2}\xi_N,$$

and therefore,

$$\widetilde{A}_N v_N = (1-\Delta)^{1/2}\xi_N + \alpha_N \phi_N,$$

where $\alpha_N = -\frac{\left\langle \phi_N, \widetilde{A}_N^{-1}(1-\Delta)^{1/2}\xi_N\right\rangle_{\mathrm{L}^2_\#}}{\left\langle \phi_N, \widetilde{A}_N^{-1}\phi_N\right\rangle_{\mathrm{L}^2_\#}} = -\frac{\left\langle (1-\Delta)^{-1/2}\phi_N, B_N\xi_N\right\rangle_{\mathrm{L}^2_\#}}{\left\langle \phi_N, \widetilde{A}_N^{-1}\phi_N\right\rangle_{\mathrm{L}^2_\#}} \in \mathbb{R}$ is characterized by the constraint $v_N \in \phi_N^\perp$. We thus obtain

$$v_N = \widetilde{A}_N^{-1}(1-\Delta)^{1/2}\xi_N - \frac{\left\langle (1-\Delta)^{-1/2}\phi_N, B_N\xi_N\right\rangle_{\mathrm{L}^2_\#}}{\left\langle \phi_N, \widetilde{A}_N^{-1}\phi_N\right\rangle_{\mathrm{L}^2_\#}} \widetilde{A}_N^{-1}\phi_N,$$

and therefore, as $v_N \in \phi_N^\perp$,

$$\begin{aligned}
&M_N^{1/2}A_N^{-1}M_N^{1/2}\xi_N - \xi_N \\
&\quad = M_N^{1/2}v_N - \xi_N \\
&\quad = \Pi_{\phi_N}^\perp (1-\Delta)^{1/2}v_N - \xi_N \\
&\quad = (1-\Delta)^{1/2}v_N - \left\langle \phi_N, (1-\Delta)^{1/2}v_N\right\rangle_{\mathrm{L}^2_\#}\phi_N - \xi_N \\
&\quad = (B_N-1)\xi_N - \langle \phi_N, B_N\xi_N\rangle_{\mathrm{L}^2_\#}\phi_N \\
&\qquad - \frac{\left\langle (1-\Delta)^{-1/2}\phi_N, B_N\xi_N\right\rangle_{\mathrm{L}^2_\#}}{\left\langle \phi_N, \widetilde{A}_N^{-1}\phi_N\right\rangle_{\mathrm{L}^2_\#}}\left((1-\Delta)^{1/2}\widetilde{A}_N^{-1}\phi_N - \left\langle \phi_N, (1-\Delta)^{1/2}\widetilde{A}_N^{-1}\phi_N\right\rangle_{\mathrm{L}^2_\#}\phi_N\right) \\
&\quad = (B_N-1)\xi_N - \langle \phi_N, B_N\xi_N\rangle_{\mathrm{L}^2_\#}\phi_N \\
&\qquad - \frac{\left\langle (1-\Delta)^{-1/2}\phi_N, B_N\xi_N\right\rangle_{\mathrm{L}^2_\#}}{\left\langle \phi_N, \widetilde{A}_N^{-1}\phi_N\right\rangle_{\mathrm{L}^2_\#}}\left(B_N(1-\Delta)^{-1/2}\phi_N - \left\langle \phi_N, B_N(1-\Delta)^{-1/2}\phi_N\right\rangle_{\mathrm{L}^2_\#}\phi_N\right) \\
&\quad = (B_N-1)\xi_N - \langle \phi_N, (B_N-1)\xi_N\rangle_{\mathrm{L}^2_\#}\phi_N \\
&\qquad - \frac{\left\langle (1-\Delta)^{-1/2}\phi_N, (B_N-1)\xi_N\right\rangle_{\mathrm{L}^2_\#}}{\left\langle \phi_N, \widetilde{A}_N^{-1}\phi_N\right\rangle_{\mathrm{L}^2_\#}}\left(B_N(1-\Delta)^{-1/2}\phi_N - \left\langle \phi_N, B_N(1-\Delta)^{-1/2}\phi_N\right\rangle_{\mathrm{L}^2_\#}\phi_N\right),
\end{aligned}$$

where we used the fact that $\xi_N \in \mathcal{X}_N^\perp$, while ϕ_N and $(1-\Delta)^{-1/2}\phi_N$ belong to $\mathcal{X}_N$. Using again (37) we obtain that for N large enough,

$$\begin{aligned}
\left\|M_N^{1/2}A_N^{-1}M_N^{1/2} - I_{\mathcal{X}_N^\perp}\right\|_{\mathcal{X}_N^\perp \to \mathrm{L}^2_\#} &= \sup_{\xi_N \in \mathcal{X}_N^\perp} \frac{\left\|M_N^{1/2}A_N^{-1}M_N^{1/2}\xi_N - \xi_N\right\|_{\mathrm{L}^2_\#}}{\|\xi_N\|_{L^2_\#}} \\
&\leqslant C_* \left\|(B_N-1)\Pi_N^\perp\right\|_{\mathrm{L}^2_\# \to \mathrm{L}^2_\#},
\end{aligned}$$

where

$$C_* := 3 + \frac{\|\phi_*\|_{\mathrm{H}^{-1}_\#}}{\left\langle \phi_*, \widetilde{A}_*^{-1}\phi_*\right\rangle_{\mathrm{L}^2_\#}} \times 2\|B_*\|_{\mathrm{L}^2_\# \to \mathrm{L}^2_\#}\|\phi_*\|_{\mathrm{H}^{-1}_\#}.$$

Finally, using (41), we have

$$\begin{aligned}
\left\|(B_N-1)\Pi_N^\perp\right\|_{\mathrm{L}^2_\# \to \mathrm{L}^2_\#} &\leqslant \left\|B_N(1-\Delta)^{-1/2}W_N\right\|_{\mathrm{L}^2_\# \to \mathrm{L}^2_\#}\left\|(1-\Delta)^{-1/2}\Pi_N^\perp\right\|_{\mathrm{L}^2_\# \to \mathrm{L}^2_\#} \\
&\leqslant \|B_N\|_{\mathrm{L}^2_\# \to \mathrm{L}^2_\#}\|W_N\|_{\mathrm{L}^\infty_\#}(1+2N)^{-1/2} \underset{N\to 0}{\longrightarrow} 0,
\end{aligned}$$

since $\|B_N\|_{\mathrm{L}^2_\# \to \mathrm{L}^2_\#}$ and $\|W_N\|_{\mathrm{L}^\infty_\#}$ are uniformly bounded in N. This concludes the proof. □

References

[1] E. Cancès, R. Chakir, and Y. Maday. Numerical Analysis of Nonlinear Eigenvalue Problems. *Journal of Scientific Computing*, 45(1):90–117, 2010.

[2] E. Cancès, G. Dusson, Y. Maday, B. Stamm, and M. Vohralik. Guaranteed and Robust a Posteriori Bounds for Laplace Eigenvalues and Eigenvectors: Conforming Approximations. *SIAM Journal on Numerical Analysis*, 55(5):2228–2254, 2017.

[3] E. Cancès, G. Dusson, Y. Maday, B. Stamm, and M. Vohralik. Guaranteed and robust a posteriori bounds for Laplace eigenvalues and eigenvectors: A unified framework. *Numerische Mathematik*, 140(4):1033–1079, 2018.

[4] E. Cancès, G. Dusson, Y. Maday, B. Stamm, and M. Vohralik. Guaranteed a posteriori bounds for eigenvalues and eigenvectors: Multiplicities and clusters. *Mathematics of Computation*, 2020.

[5] E. Cancès, G. Dusson, Y. Maday, B. Stamm, and M. Vohralik. Post-processing of the planewave approximation of Schrödinger equations. Part I: Linear operators. *IMA Journal of Numerical Analysis*, (draa044), 2020.

[6] E. Cancès, V. Ehrlacher, D. Gontier, A. Levitt, and D. Lombardi. Numerical quadrature in the Brillouin zone for periodic Schrödinger operators. *Numerische Mathematik*, 144(3):479–526, 2020.

[7] E. Cancès, G. Kemlin, and A. Levitt. Convergence Analysis of Direct Minimization and Self-Consistent Iterations. *SIAM Journal on Matrix Analysis and Applications*, 42(1):243–274, 2021.

[8] C. Carstensen and J. Gedicke. Guaranteed lower bounds for eigenvalues. *Mathematics of Computation*, 83(290):2605–2629, 2014.

[9] G. Dusson. Post-processing of the plane-wave approximation of Schrödinger equations. Part II: Kohn–Sham models. *IMA Journal of Numerical Analysis*, (draa052), 2020.

[10] G. Dusson and Y. Maday. A posteriori analysis of a nonlinear Gross–Pitaevskii-type eigenvalue problem. *IMA Journal of Numerical Analysis*, 37(1):94–137, 2017.

[11] G. Dusson, I. Sigal, and B. Stamm. Analysis of the Feshbach-Schur method for the planewave discretizations of Schrödinger operators. *arXiv:2008.10871 [cs, math]*, 2020.

[12] S. Goedecker, M. Teter, and J. Hutter. Separable dual-space Gaussian pseudopotentials. *Physical Review B*, 54(3):1703, 1996.

[13] D. R. Hartree. The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part I. Theory and Methods. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24(1):89–110, 1928.

[14] C. Hartwigsen, S. Goedecker, and J. Hutter. Relativistic separable dual-space Gaussian pseudopotentials from H to Rn. *Physical Review B*, 58(7):3641–3662, 1998.

[15] M. F. Herbst and A. Levitt. A robust and efficient line search for self-consistent field iterations. *Journal of Computational Physics*, page 111127, Mar. 2022.

[16] M. F. Herbst and A. Levitt. A robust and efficient line search for self-consistent field iterations. *Journal of Computational Physics*, 459:111127, 2022.

[17] M. F. Herbst, A. Levitt, and E. Cancès. A posteriori error estimation for the non-self-consistent Kohn–Sham equations. *Faraday Discussions*, 224:227–246, 2020.

[18] M. F. Herbst, A. Levitt, and E. Cancès. DFTK: A Julian approach for simulating electrons in solids. *Proceedings of the JuliaCon Conferences*, 3(26):69, 2021.

[19] L. Kleinman and D. M. Bylander. Efficacious Form for Model Pseudopotentials. *Physical Review Letters*, 48(20):1425–1428, 1982.

[20] A. V. Knyazev. Toward the Optimal Preconditioned Eigensolver: Locally Optimal Block Preconditioned Conjugate Gradient Method. *SIAM Journal on Scientific Computing*, 23(2):517–541, 2001.

[21] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A):A1133–A1138, 1965.

[22] X. Liu. A framework of verified eigenvalue bounds for self-adjoint differential operators. *Applied Mathematics and Computation*, 267:341–355, 2015.

[23] R. M. Martin. *Electronic Structure: Basic Theory and Practical Methods*. Cambridge University Press, first edition, 2004.

[24] H. J. Monkhorst and J. D. Pack. Special points for Brillouin-zone integrations. *Physical Review B*, 13(12):5188–5192, 1976.

[25] M. T. Nakao, M. Plum, and Y. Watanabe. *Numerical Verification Methods and Computer-Assisted Proofs for Partial Differential Equations*. Number 53 in Springer Series in Computational Mathematics. Springer, Singapore, 2019.

[26] M. C. Payne, M. P. Teter, D. C. Allan, T. A. Arias, and J. D. Joannopoulos. Iterative minimization techniques for ab initio total-energy calculations: Molecular dynamics and conjugate gradients. *Reviews of Modern Physics*, 64(4):1045–1097, 1992.

[27] J. P. Perdew, K. Burke, and M. Ernzerhof. Generalized Gradient Approximation Made Simple. *Physical Review Letters*, 77(18):3865–3868, 1996.

[28] K. Persson. Materials data on TiO2 (SG:136) by materials project, 2014.

[29] L. P. Pitaevskii and S. Stringari. *Bose-Einstein Condensation*. International Series of Monographs on Physics. Oxford University Press, Oxford, New York, 2003.

[30] J. Revels, M. Lubin, and T. Papamarkou. Forward-Mode Automatic Differentiation in Julia. *ArXiv*, 2016.

[31] A. Schmidt, D. Wittwar, and B. Haasdonk. Rigorous and effective a-posteriori error bounds for nonlinear problems—application to RB methods. *Advances in Computational Mathematics*, 46(2):32, 2020.

[32] R. Verfürth. A posteriori error estimation and adaptive mesh-refinement techniques. *Journal of Computational and Applied Mathematics*, 50(1-3):67–83, 1994.