Post-processing of the planewave approximation of Schrödinger equations. Part I: linear operators

Eric Cancès, Geneviève Dusson, Yvon Maday, Benjamin Stamm, Martin Vohralík

▶ **To cite this version:**

Eric Cancès, Geneviève Dusson, Yvon Maday, Benjamin Stamm, Martin Vohralík. Post-processing of the planewave approximation of Schrödinger equations. Part I: linear operators. 2018. hal-01908039v2

HAL Id: hal-01908039
https://hal.science/hal-01908039v2

Preprint submitted on 20 Nov 2018 (v2), last revised 17 Mar 2020 (v4)

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.

Post-processing of the planewave approximation of Schrödinger equations. Part I: linear operators

Eric Cancès
CERMICS, Ecole des Ponts, Université Paris-Est, 6 & 8 Av. Pascal, 77455 Marne-la-Vallée, France and
Inria, 2 Rue Simone Iff, 75589 Paris, France

Geneviève Dusson*
Mathematics Institute, University of Warwick, Coventry CV47AL, UK

Yvon Maday
Sorbonne Université, Université Paris-Diderot SPC, CNRS, Laboratoire Jacques-Louis Lions (LJLL), F-75005 Paris, France, and
Institut Universitaire de France, 75005 Paris, France

Benjamin Stamm
Center for Computational Engineering Science, RWTH Aachen University, Aachen, Germany

and

Martin Vohralík
Inria, 2 Rue Simone Iff, 75589 Paris, France and
CERMICS, Ecole des Ponts, Université Paris-Est, 6 & 8 Av. Pascal, 77455 Marne-la-Vallée, France

November 12, 2018

Abstract

In this article, we prove *a priori* error estimates for the perturbation-based post-processing of the plane-wave approximation of Schrödinger equations introduced and tested numerically in previous works [6, 7]. We consider here a Schrödinger operator $\mathcal{H} = -\frac{1}{2}\Delta + \mathcal{V}$ on $L^2(\Omega)$, where Ω is a cubic box with periodic boundary conditions, and where $\mathcal{V}$ is a multiplicative operator by a regular-enough function $\mathcal{V}$. The quantities of interest are, on the one hand, the ground-state energy defined as the sum of the lowest N eigenvalues of $\mathcal{H}$, and, on the other hand, the ground-state density matrix, that is the spectral projector on the vector space spanned by the associated eigenvectors. Such a problem is central in first-principle molecular simulation, since it corresponds to the so-called linear subproblem in Kohn–Sham density functional theory (DFT). Interpreting the exact eigenpairs of $\mathcal{H}$ as perturbations of the numerical eigenpairs obtained by a variational approximation in a plane-wave (*i.e.* Fourier) basis, we compute first-order corrections for the eigenfunctions, which are turned into corrections on the ground-state density matrix. This allows us to increase the accuracy of both the ground-state energy and the ground-state density matrix at a low computational extra-cost. Indeed, the computation of the corrections only

*Corresponding author. Email: g.dusson@warwick.ac.uk

requires the computation of the residual of the solution in a larger plane-wave basis and two Fast Fourier Transforms per eigenvalue.

1 Introduction

First-principle molecular simulation is a major tool to predict the properties of matter from the atomic to the macroscopic scales. It is widely used in different fields such as chemistry, condensed matter physics, or materials science. As a main advantage, it requires no empirical parameter except a few fundamental constants of physics (the reduced Planck constant $\hbar$, the electron mass m_e, the elementary charge e, the dielectric permittivity of the vacuum ε_0, the Boltzmann constant k_B), as well as the masses and atomic numbers of the nuclei contained in the system under consideration.

At this scale, matter is described as a system of nuclei and electrons, whose dynamics is modeled by a time-dependent many-body Schrödinger equation. This equation, which is an evolution partial differential equation on a $3(M+N)$-dimensional space, where M is the number of nuclei and N the number of electrons, is way too costly to be solved in practice when $(M+N)$ exceeds 2 or 3. Hence different approximations have to be resorted to, starting with considering a simpler model. First, in almost all molecular simulations, the nuclei, which are thousands times heavier than electrons, are considered as point-like classical particles, and the electrons are supposed to be, at each time t, in their ground-state. This is called the Born–Oppenheimer approximation [1].

Many different approaches have then been proposed to compute the electronic ground-state. The most popular ones can be classified into three main classes:

- wavefunction methods, among which the Hartree–Fock, post Hartree–Fock, and multi-reference methods (see [11], and [4] for a mathematical introduction);
- density functional theory (DFT) methods, consisting of orbital-free and Kohn–Sham models [9, 14];
- quantum Monte Carlo methods [15, 16].

The Kohn–Sham models are the most widely used in physics and chemistry, as they provide a good compromise between accuracy and computational cost, being posed only in a 3-dimensional space. In condensed matter physics and materials science, most Kohn–Sham calculations are performed in a rectangular box Ω, called a supercell, with periodic boundary conditions (Born–von Karman PBC). The most common method to discretize the Kohn–Sham equations then is to use a variational approximation in a plane-wave (Fourier) basis. For very large systems, unfortunately, using a fine discretization basis is too expensive, while using a coarse discretization basis leads to insufficiently accurate results.

In order to limit the computational cost of the method while preserving the quality of the numerical results, several post-processing methods have been proposed. Usually, the approach is to perform a full computation in a coarse basis, for which the computational cost is not excessively high, and then to make some not-too-expensive computation in a finer basis leading to a substantial improvement in accuracy.

In general, this approach falls into the category of two-grid methods, which have been applied to the case of a linear eigenvalue problem in [19]. It has then been extended to a class of elliptic nonlinear eigenvalue problems in [2]. In that paper, the *nonlinear* eigenvalue problem is first solved on a coarse grid, and then a *linear* eigenvalue problem or a boundary problem is solved on a finer grid. Other post-processing methods have been proposed for nonlinear eigenvalue problems, for example in [17] for the Hartree–Fock problem (see also the references therein).

In [6, 7], we introduced a new post-processing method for periodic Kohn–Sham calculations in a plane-wave basis, leading to a significant gain in accuracy at a very limited computational extra-cost. This approach is based on the Rayleigh–Schrödinger perturbation method, considering the exact Kohn–Sham ground-state as a perturbation of the approximate ground-state computed in a finite basis set. Theoretical estimates in the asymptotic regime for the energy were announced and illustrated by numerical simulations. These

simulations showed that the accuracy on the ground-state energy could be improved by a factor of 10 to 100 at a very limited extra-cost (about 3%).

In this contribution, we focus on the linear subproblem of the Kohn–Sham model. We apply the post-processing of the plane-wave approximation of Schrödinger equations introduced in [6, 7] and present the proof of the improved accuracy. The linear subproblem consists in computing the rank–N ground-state density matrix γ_0 of a linear Schrödinger operator $\mathcal{H} = -\frac{1}{2}\Delta + \mathcal{V}$, acting on the space $L^2_\#(\Omega)$ of real-valued square-integrable periodic functions on $\mathbb{R}^3$ with Ω as a periodic cell. For $\mathcal{V} \in L^2_\#(\Omega)$, the Hamiltonian $\mathcal{H}$ is diagonalizable in an orthonormal basis and its eigenvalues $\lambda_1^0 \leq \lambda_2^0 \leq \cdots$ (counting multiplicities) form a non-decreasing sequence of real numbers that tends to $+\infty$. Denoting by $(\phi_i^0)_{i\geq 1}$ an orthonormal basis of associated eigenvectors, and assuming that there is a gap $g := \lambda_{N+1}^0 - \lambda_N^0 > 0$ between the N^{th} and the $(N+1)^{\text{st}}$ eigenvalue of $\mathcal{H}$, the rank–N ground-state density matrix is defined, using Dirac's bra-ket notation, as

$$\gamma_0 = \sum_{i=1}^{N} |\phi_i^0\rangle\langle\phi_i^0|.$$

It is therefore the orthogonal projector on the N-dimensional vector subspace of $L^2_\#(\Omega)$ spanned by the eigenvectors associated with the lowest N eigenvalues of $\mathcal{H}$. The ground-state energy is then defined as the scalar quantity

$$\mathcal{E}_0 = \sum_{i=1}^{N} \lambda_i^0.$$

Note that $\mathcal{E}_0$ is not equal to the Kohn–Sham ground-state energy when $\mathcal{H}$ is the Kohn–Sham Hamiltonian due to nonlinear effects, called double-counting in the chemistry and physics literatures.

Our main result is summarized in Theorem 4.1. We show that, in the asymptotic regime where the discretization space is large enough, the convergence rates of both the post-processed ground-state density matrix (built from the post-processed eigenvectors defined in [6]) and the post-processed ground-state energy are improved. Note that we do not make any non-degeneracy assumption on the lowest N eigenvalues of $\mathcal{H}$; only the presence of a positive gap between the N^{th} and the $(N+1)^{\text{st}}$ eigenvalues of $\mathcal{H}$ is required. As in [6, 7], our approach strongly relies on the fact that, in plane-wave approximations, the kinetic energy operator $-\frac{1}{2}\Delta$, which is the leading term in the Hamiltonian $\mathcal{H}$, commutes with the orthogonal projector on the discretization space.

This article is organized as follows. In Section 2.1, we present in detail the linear subproblem of the Kohn–Sham model, as well as the characterization of γ_0 as the unique solution to some constrained optimization problem, and some other useful classical results. In Section 2.3, we describe the plane-wave discretization of this optimization problem. In Section 2.4, we translate the *a priori* error analysis results of [3] into the density matrix formalism. Our post-processing method based on Rayleigh–Schrödinger perturbation theory is described in Section 3. In Section 4.1, we present the main results of this paper, *i.e.* an improved convergence rate on the post-processed ground-state density matrix and energy. The proofs are given in Section 4.2. Some numerical simulations are presented in Section 5. The case of the nonlinear Kohn–Sham model will be dealt with in a forthcoming paper [10].

2 Post-processing for the Kohn–Sham linear subproblem

In order to simplify the notation, we consider a cubic lattice $\mathcal{R} = L\mathbb{Z}^3$ ($L > 0$) corresponding to the supercell $\Omega = [0, L)^3$, but all our arguments straightforwardly apply to the general case of a lattice with lower or no point symmetry. For $1 \leq p \leq \infty$ and $s \in \mathbb{R}_+$, we denote by

$$L^p_\#(\Omega) := \left\{u \in L^p_{\text{loc}}(\mathbb{R}^3, \mathbb{R}) \mid u \text{ is } \mathcal{R}\text{-periodic}\right\},$$
$$H^s_\#(\Omega) := \left\{u \in H^s_{\text{loc}}(\mathbb{R}^3, \mathbb{R}) \mid u \text{ is } \mathcal{R}\text{-periodic}\right\},$$

the spaces of real-valued $\mathcal{R}$-periodic L^p and H^s functions, and by $\mathcal{L}(L^2_\#)$ the vector space of the bounded linear operators on $L^2_\#(\Omega)$.

2.1 Problem setting

Let $N \in \mathbb{N}^*$ and $\mathcal{V} \in L^2_\#(\Omega)$. In Kohn–Sham models, N is the number of electrons (or of electron pairs in closed-shell models) in the simulation cell, and $\mathcal{V}$ is an approximation of the Kohn–Sham effective potential. Let $\mathcal{H}$ be the operator on $L^2_\#(\Omega)$ with domain $H^2_\#(\Omega)$ defined by

$$\forall u \in H^2_\#(\Omega), \quad \mathcal{H}\, u = -\frac{1}{2}\Delta u + \mathcal{V} u.$$

It is well-known that the operator $\mathcal{H}$ is self-adjoint, bounded below, with compact resolvent. It can therefore be diagonalized in an orthonormal basis: there exists a non-decreasing sequence $(\lambda_i^0)_{i\geq 1}$ of real numbers and an orthonormal basis $(\phi_i^0)_{i\geq 1}$ of $L^2_\#(\Omega)$ consisting of functions of $H^2_\#(\Omega)$ such that

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

We denote by $\Phi^0 = (\phi_1^0, \cdots, \phi_N^0)^T$ and $\Lambda^0 = \mathrm{diag}(\lambda_1^0, \cdots, \lambda_N^0)$.

A key assumption for our analysis is the following:

Assumption 2.1. *There is a gap between the N^{th} and the $(N+1)^{\text{st}}$ eigenvalues of $\mathcal{H}$, i.e.*

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

We denote by $\epsilon_{\mathrm{F}} := \frac{\lambda_N^0 + \lambda_{N+1}^0}{2}$ the Fermi level. Note that, in this setting, any real number in the range $(\lambda_N^0, \lambda_{N+1}^0)$ is an admissible Fermi level.

As already mentioned in the introduction, the purpose of the linear subproblem is to compute two quantities of interest:

1. the ground-state density matrix

$$\boxed{\gamma_0 := \mathbb{1}_{(-\infty, \epsilon_{\mathrm{F}}]}(\mathcal{H}) = \sum_{i=1}^{N} |\phi_i^0\rangle\langle\phi_i^0|;}$$

2. the ground-state energy

$$\boxed{\mathcal{E}_0 := \mathrm{Tr}\,(\mathcal{H}\,\gamma_0) = \sum_{i=1}^{N} \lambda_i^0,}$$

where Tr denotes the trace, and will be properly introduced in Section 2.2.

The linear subproblem can be formulated as a variational problem in several ways. First, introducing the quadratic form

$$H^1_\#(\Omega) \ni \psi \mapsto \langle\psi|\mathcal{H}\,|\psi\rangle := \frac{1}{2}\int_\Omega |\nabla\psi|^2 + \int_\Omega \mathcal{V}\,|\psi|^2 \in \mathbb{R}$$

associated with $\mathcal{H}$, the energy functional $\mathcal{E}$ defined by

$$\forall \Psi = (\psi_1, \ldots, \psi_N)^T \in \left[H^1_\#(\Omega)\right]^N, \quad \mathcal{E}(\Psi) := \sum_{i=1}^{N} \langle\psi_i|\mathcal{H}\,|\psi_i\rangle = \sum_{i=1}^{N} \left(\frac{1}{2}\int_\Omega |\nabla\psi_i|^2 + \int_\Omega \mathcal{V}\,|\psi_i|^2\right), \tag{1}$$

and the (infinite-dimensional) Stiefel manifold

$$\mathcal{M} = \left\{ \Psi = (\psi_1, \ldots, \psi_N)^T \in \left[H^1_\#(\Omega)\right]^N \;\middle|\; \forall i,j = 1, \ldots, N, \; \int_\Omega \psi_i\psi_j = \delta_{ij} \right\}, \tag{2}$$

we have

$$\mathcal{E}_0 = \inf \left\{ \mathcal{E}(\Psi),\ \Psi \in \mathcal{M} \right\}. \tag{3}$$

Besides, $\Phi^0 = (\phi_1^0, \dots, \phi_N^0)^T$ is a minimizer of (3). Note that Φ^0 is not the unique minimizer of (3). Indeed, denoting by $O(N) := \left\{ U \in \mathbb{R}^{N \times N} \mid U^T U = 1_N \right\}$ the orthogonal group in dimension N (1_N is the identity matrix of rank N), we have

$$\forall \Psi \in \mathcal{M}, \quad \forall U \in O(N), \quad U\Psi \in \mathcal{M}, \quad \text{and} \quad \mathcal{E}(U\Psi) = \mathcal{E}(\Psi). \tag{4}$$

Therefore, $U\Phi^0$ is a minimizer of (3) for all $U \in O(N)$. In fact, under Assumption 2.1, the set of minimizers of (3) is exactly equal to $O(N)\Phi^0$. For the sake of completeness, let us recall the proof of this elementary, but key, property. Let $\Psi = (\psi_1, \cdots, \psi_N)^T$ be a critical point of (3). The first-order optimality conditions satisfied by Ψ read

$$\forall i, j = 1, \dots, N, \quad \mathcal{H}\,\psi_i = \sum_{j=1}^{N} \lambda_{ij} \psi_j, \quad \int_\Omega \psi_i \psi_i = \delta_{ij}.$$

The $N \times N$ symmetric matrix $\Lambda = (\lambda_{ij})_{i,j=1,\dots,N}$ is the Lagrange multiplier of the matrix constraint $\int_\Omega \psi_i \psi_j = \delta_{ij}$. It is not diagonal in general. On the other hand, since it is symmetric, there exists $U \in O(N)$ such that $U\Lambda U^T = \mathrm{diag}(\lambda_1, \cdots, \lambda_N)$ with $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N$. Then, $\Phi = (\phi_1, \cdots, \phi_N)^T := U\Psi$ also is a critical point of (3) with the same energy as Ψ, and we have

$$\forall 1 \leq i, j \leq N, \quad \mathcal{H}\,\phi_i = \lambda_i \phi_i, \quad \int_\Omega \phi_i \phi_j = \delta_{ij} \quad \text{and} \quad \mathcal{E}(\Phi) = \mathcal{E}(\Psi) = \sum_{i=1}^{N} \lambda_i.$$

For Ψ to be a minimizer of (3), we must have $\lambda_i = \lambda_i^0$ for all $1 \leq i \leq N$. Under Assumption 2.1, we have in addition $\mathrm{Span}(\psi_1, \cdots, \psi_N) = \mathrm{Span}(\phi_1, \cdots, \phi_N) = \mathrm{Span}(\phi_1^0, \cdots, \phi_N^0) = \mathrm{Ran}(\gamma_0)$. Therefore, there exists $U \in O(N)$ such that $\Psi = U\Phi_0$.

To get rid of the gauge invariance (4), it is convenient to reformulate problem (3) in terms of density matrices. Introducing the (infinite-dimensional) Grassmann manifold

$$\Upsilon = \left\{ \gamma \in \mathcal{L}(L^2_\#) \;\middle|\; \gamma^* = \gamma,\ \gamma^2 = \gamma,\ \mathrm{Tr}\,(\gamma) = N,\ \mathrm{Tr}\,(-\Delta\gamma) < \infty \right\}, \tag{5}$$

its convex hull

$$\mathcal{K} = \left\{ \gamma \in \mathcal{L}(L^2_\#) \;\middle|\; \gamma^* = \gamma,\ 0 \leq \gamma \leq 1,\ \mathrm{Tr}\,(\gamma) = N,\ \mathrm{Tr}\,(-\Delta\gamma) < \infty \right\}, \tag{6}$$

and the energy functional E defined on $\mathcal{K}$ by

$$\forall \gamma \in \mathcal{K}, \quad E(\gamma) = \mathrm{Tr}\,(\mathcal{H}\,\gamma), \tag{7}$$

it holds

$$\mathcal{E}_0 = \inf \left\{ E(\gamma),\ \gamma \in \Upsilon \right\}, \tag{8}$$

and

$$\mathcal{E}_0 = \inf \left\{ E(\gamma),\ \gamma \in \mathcal{K} \right\}. \tag{9}$$

Besides, under Assumption 2.1, γ_0 is the unique minimizer of both (8) and (9). Here γ^* denotes the adjoint of γ, $0 \leq \gamma \leq 1$ means $\forall u \in L^2_\#(\Omega)$, $0 \leq \langle u | \gamma\, u \rangle \leq \|u\|^2_{L^2_\#}$. The precise meaning of the terms $\mathrm{Tr}\,(-\Delta\gamma)$ and $\mathrm{Tr}\,(\mathcal{H}\,\gamma)$, as well as the proof of the fact that γ_0 is the unique minimizer of (8) and (9), will be given in the next section.

2.2 Functional setting

We denote by $\|\cdot\|$ the operator norm on $\mathcal{L}(L^2_\#)$, the space of bounded linear operators on $L^2_\#(\Omega)$. We also need to introduce the Banach space $\mathfrak{S}_1(L^2_\#)$ of trace-class operators on $L^2_\#(\Omega)$ and the Hilbert space $\mathfrak{S}_2(L^2_\#)$ of Hilbert–Schmidt operators on $L^2_\#(\Omega)$, respectively endowed with the norm defined by $\|A\|_{\mathfrak{S}_1(L^2_\#)} := \mathrm{Tr}\,(|A|) = \mathrm{Tr}\,(\sqrt{A^*A})$ and the inner product defined by $(A,B)_{\mathfrak{S}_2(L^2_\#)} := \mathrm{Tr}\,(A^*B)$. We refer to [18, Chapter VI] for an introduction to trace-class and Hilbert–Schmidt operators. Let us just recall here the properties which will be used in the sequel:

- for any orthonormal basis $(e_n)_{n\in\mathbb{N}}$ of $L^2_\#(\Omega)$, we have

$$\forall A \in \mathfrak{S}_1(L^2_\#), \quad \mathrm{Tr}\,(A) = \sum_{n\in\mathbb{N}} \langle e_n | A e_n \rangle,$$

$$\forall A \in \mathfrak{S}_2(L^2_\#), \quad \|A\|_{\mathfrak{S}_2(L^2_\#)} = \mathrm{Tr}\,(A^*A)^{1/2} = \left(\sum_{n\in\mathbb{N}} \|Ae_n\|^2_{L^2_\#} \right)^{1/2}.$$

  If $A \in \mathcal{L}(L^2_\#)$ is a *positive* operator, that is if for all $u \in L^2_\#(\Omega)$, there holds that $\langle u|Au\rangle \geq 0$, then the value of the sum

$$\mathrm{Tr}\,(A) := \sum_{n\in\mathbb{N}} \langle e_n | A e_n \rangle \in \mathbb{R}_+ \cup \{+\infty\}$$

  is independent of the choice of the orthonormal basis $(e_n)_{n\in\mathbb{N}}$. If $A \in \mathcal{L}(L^2_\#)$ is *positive and self-adjoint*, then $A \in \mathfrak{S}_1(L^2_\#)$ if and only if $\mathrm{Tr}\,(A) < \infty$;

- $\mathfrak{S}_1(L^2_\#) \subset \mathfrak{S}_2(L^2_\#) \subset \mathcal{L}(L^2_\#)$ and for all $A \in \mathfrak{S}_1(L^2_\#)$,

$$\|A\| \leq \|A\|_{\mathfrak{S}_2(L^2_\#)} \leq \|A\|_{\mathfrak{S}_1(L^2_\#)}; \tag{10}$$

- for all $A \in \mathfrak{S}_1(L^2_\#)$ and $B \in \mathcal{L}(L^2_\#)$, we have $AB \in \mathfrak{S}_1(L^2_\#)$, $BA \in \mathfrak{S}_1(L^2_\#)$,

$$\mathrm{Tr}\,(AB) = \mathrm{Tr}\,(BA), \quad \|AB\|_{\mathfrak{S}_1(L^2_\#)} \leq \|B\|\,\|A\|_{\mathfrak{S}_1(L^2_\#)}, \quad \|BA\|_{\mathfrak{S}_1(L^2_\#)} \leq \|B\|\,\|A\|_{\mathfrak{S}_1(L^2_\#)}; \tag{11}$$

- for all $A \in \mathfrak{S}_1(L^2_\#)$, there exists a unique function $\rho_A \in L^1_\#(\Omega)$, called the density associated with the operator A, such that for all $V \in L^\infty_\#(\Omega)$,

$$\mathrm{Tr}\,(AV) = \int_\Omega \rho_A V,$$

  where on the left–hand side of the above equality, $V \in \mathcal{L}(L^2_\#)$ denotes the multiplication operator by the function V;

- for all $A \in \mathfrak{S}_2(L^2_\#)$ and $B \in \mathcal{L}(L^2_\#)$, we have $AB \in \mathfrak{S}_2(L^2_\#)$, $BA \in \mathfrak{S}_2(L^2_\#)$,

$$\|AB\|_{\mathfrak{S}_2(L^2_\#)} \leq \|B\|\,\|A\|_{\mathfrak{S}_2(L^2_\#)}, \quad \|BA\|_{\mathfrak{S}_2(L^2_\#)} \leq \|B\|\,\|A\|_{\mathfrak{S}_2(L^2_\#)}; \tag{12}$$

- for all $A \in \mathfrak{S}_2(L^2_\#)$ and $B \in \mathfrak{S}_2(L^2_\#)$, $AB \in \mathfrak{S}_1(L^2_\#)$, $BA \in \mathfrak{S}_1(L^2_\#)$,

$$\mathrm{Tr}\,(AB) = \mathrm{Tr}\,(BA) \leq \|A\|_{\mathfrak{S}_2(L^2_\#)} \|B\|_{\mathfrak{S}_2(L^2_\#)}. \tag{13}$$

We set

$$\forall \Psi = (\psi_1, \cdots, \psi_N)^T \in \left[L^2_\#(\Omega)\right]^N, \quad \|\Psi\|_{L^2_\#} := \left(\sum_{i=1}^N \|\psi_i\|^2_{L^2_\#}\right)^{1/2},$$

$$\forall \Psi = (\psi_1, \cdots, \psi_N)^T \in \left[H^1_\#(\Omega)\right]^N, \quad \|\Psi\|_{H^1_\#} := \|(1-\Delta)^{1/2}\Psi\|_{L^2_\#} = \left(\sum_{i=1}^N \|\psi_i\|^2_{H^1_\#}\right)^{1/2},$$

and more generally, for any operator A on $L^2_\#(\Omega)$ with domain $D(A)$,

$$\forall \Psi \in [D(A)]^N, \quad \|A\Psi\|_{L^2_\#} := \left(\sum_{i=1}^N \|A\psi_i\|^2_{L^2_\#}\right)^{1/2}.$$

Let $\mathcal{R}^* = \frac{2\pi}{L}\mathbb{Z}^3$ be the dual lattice of the periodic lattice $\mathcal{R} = L\mathbb{Z}^3$. For $\mathbf{k} \in \mathcal{R}^*$, we denote by $e_\mathbf{k}$ the plane-wave with wavevector $\mathbf{k}$, defined by

$$\begin{aligned} e_\mathbf{k} : \quad & \mathbb{R}^3 \to \mathbb{C} \\ & \mathbf{x} \mapsto |\Omega|^{-1/2} e^{i\mathbf{k}\cdot\mathbf{x}}, \end{aligned}$$

where $|\Omega| = L^3$. The family $(e_\mathbf{k})_{\mathbf{k}\in\mathcal{R}^*}$ forms an orthonormal basis of the complex Hilbert space

$$L^2_\#(\Omega, \mathbb{C}) := \left\{u \in L^2_{\rm loc}(\mathbb{R}^3, \mathbb{C}) \mid u \text{ is } \mathcal{R}\text{-periodic}\right\},$$

endowed with the scalar product

$$\forall u, v \in L^2_\#(\Omega, \mathbb{C}), \quad \langle u|v\rangle = \int_\Omega \overline{u(\mathbf{r})}\, v(\mathbf{r})\, d\mathbf{r},$$

where $\overline{u(\mathbf{r})}$ denotes the complex conjugate of $u(\mathbf{r})$, and for all $v \in L^2_\#(\Omega, \mathbb{C})$,

$$v(\mathbf{r}) = \sum_{\mathbf{k}\in\mathcal{R}^*} \widehat{v}_\mathbf{k}\, e_\mathbf{k}(\mathbf{r}) \qquad \text{with} \qquad \widehat{v}_\mathbf{k} = \langle e_\mathbf{k}|v\rangle = |\Omega|^{-1/2}\int_\Omega v(\mathbf{r}) e^{-i\mathbf{k}\cdot\mathbf{r}}\, d\mathbf{r}.$$

Recall that the periodic Sobolev spaces $H^s_\#(\Omega)$ can be characterized in a simple way using Fourier series: for $s \in \mathbb{R}$, we have

$$H^s_\#(\Omega) := \left\{ v = \sum_{\mathbf{k}\in\mathcal{R}^*} \widehat{v}_\mathbf{k} e_\mathbf{k} \;\middle|\; \forall\, \mathbf{k}, \quad \widehat{v}_{-\mathbf{k}} = \overline{\widehat{v}_\mathbf{k}}, \quad \|v\|^2_{H^s_\#} := \sum_{\mathbf{k}\in\mathcal{R}^*} (1+|\mathbf{k}|^2)^s |\widehat{v}_\mathbf{k}|^2 < \infty \right\},$$

where the $H^s_\#$ inner product is defined by

$$\forall u, v \in H^s_\#(\Omega), \quad (u, v)_{H^s_\#} := \sum_{\mathbf{k}\in\mathcal{R}^*} (1+|\mathbf{k}|^2)^s \overline{u_\mathbf{k}} v_\mathbf{k}.$$

Let us now clarify the meaning of the terms $\mathrm{Tr}\,(-\Delta\gamma)$ and $\mathrm{Tr}\,(\mathcal{H}\,\gamma)$ appearing in (5)–(7). Let $\gamma \in \mathcal{L}(L^2_\#)$ be self-adjoint and positive. Since $|\nabla|$ (*i.e.* the multiplication operator by $|\mathbf{k}|$ in Fourier representation) is a bounded linear operator from $H^s_\#(\Omega)$ to $H^{s-1}_\#(\Omega)$, $|\nabla|\gamma|\nabla|$ defines a bounded linear operator from $H^1_\#(\Omega)$ to $H^{-1}_\#(\Omega)$. If in addition, $\mathrm{Ran}(|\nabla|\gamma|\nabla|) \subset L^2_\#(\Omega)$ and

$$\exists C \in \mathbb{R}_+ \quad \text{such that} \quad \forall u \in H^1_\#(\Omega), \quad \||\nabla|\gamma|\nabla|u\|_{L^2_\#} \le C\|u\|_{L^2_\#},$$

then $|\nabla|\gamma|\nabla|$ can be uniquely extended to a bounded, self-adjoint, positive operator on $L^2_\#(\Omega)$, also denoted by $|\nabla|\gamma|\nabla|$ for simplicity. In this case, $\mathrm{Tr}\,(|\nabla|\gamma|\nabla|)$ is well-defined in $\mathbb{R}_+ \cup \{+\infty\}$. In view of the fact that $-\Delta = |\nabla|^2$, the notation

$$\mathrm{Tr}\,(-\Delta\gamma) := \mathrm{Tr}\,(|\nabla|\gamma|\nabla|)$$

is commonly used in the mathematical physics literature. Let us emphasize that $\mathrm{Tr}\,(-\Delta\gamma) < \infty$ only means that $\mathrm{Tr}\,(|\nabla|\gamma|\nabla|) < \infty$; in particular, it does *not* imply that the operator $-\Delta\gamma$ is in $\mathfrak{S}_1(L^2_\#)$.

It follows from the Hoffmann–Ostenhof inequality [12] that for all $\gamma \in \mathcal{K}$, $\rho_\gamma \geq 0$ and $\sqrt{\rho_\gamma} \in H^1_\#(\Omega) \hookrightarrow L^6_\#(\Omega)$. The real number $\mathrm{Tr}\,(\mathcal{H}\,\gamma)$ can therefore be defined for all $\gamma \in \mathcal{K}$ as

$$\mathrm{Tr}\,(\mathcal{H}\,\gamma) := \frac{1}{2}\mathrm{Tr}\,(-\Delta\gamma) + \int_\Omega \rho_\gamma V.$$

It is known in addition (see e.g. [5]) that, under Assumption 2.1, there exist $0 < c \leq C < \infty$ such that

$$c(1-\Delta) \leq |\mathcal{H} - \epsilon_{\mathrm{F}}| \leq C(1-\Delta), \tag{14}$$

where $|\mathcal{H} - \epsilon_{\mathrm{F}}| = -\gamma_0(\mathcal{H} - \epsilon_F)\gamma_0 + (1-\gamma_0)(\mathcal{H} - \epsilon_F)(1-\gamma_0)$ is defined by functional calculus for self-adjoint operators, and

$$\forall \gamma \in \mathcal{K}, \quad \mathrm{Tr}\,(\mathcal{H}\,\gamma) - \mathrm{Tr}\,(\mathcal{H}\,\gamma_0) = \||\mathcal{H} - \epsilon_F|^{1/2}(\gamma - \gamma_0)\|^2_{\mathfrak{S}_2(L^2_\#)}. \tag{15}$$

We deduce from (14) and (15) that there exist two constants $0 < c \leq C < \infty$ such that

$$\forall \gamma \in \mathcal{K}, \quad c\|(1-\Delta)^{1/2}(\gamma - \gamma_0)\|^2_{\mathfrak{S}_2(L^2_\#)} \leq E(\gamma) - \mathcal{E}_0 \leq C\|(1-\Delta)^{1/2}(\gamma - \gamma_0)\|^2_{\mathfrak{S}_2(L^2_\#)}. \tag{16}$$

This implies in particular that γ_0 is the unique minimizer of (8) and (9).

Note that for all $\gamma \in \Upsilon$, as $\gamma^2 = \gamma$ and by the cyclicity of the trace, there also holds $\mathrm{Tr}\,(\mathcal{H}\,\gamma) = \mathrm{Tr}\,(\gamma\mathcal{H}\,) = \mathrm{Tr}\,(\gamma\mathcal{H}\,\gamma)$.

2.3 Discretization

In order to solve problem (3) numerically, we use a plane-wave discretization. For each $\mathbf{k} \in \mathcal{R}^*$, the kinetic energy of the plane-wave $e_{\mathbf{k}}$ is given by $\frac{1}{2}|\mathbf{k}|^2$, where $|\cdot|$ denotes the Euclidean norm. To construct a discretization space, we introduce some energy cut-off $E_{\mathrm{c}} > 0$ and consider all plane-waves whose kinetic energy is smaller than E_{c}, *i.e.* $|\mathbf{k}| \leq \sqrt{2E_{\mathrm{c}}}$. For each cut-off energy E_{c}, we set $N_{\mathrm{c}} = \sqrt{\frac{E_{\mathrm{c}}}{2}}\frac{L}{\pi}$ and

$$X_{N_{\mathrm{c}}} := \left\{ \sum_{\mathbf{k}\in\mathcal{R}^*, |\mathbf{k}| \leq \frac{2\pi}{L}N_{\mathrm{c}}} \widehat{v}_{\mathbf{k}}\, e_{\mathbf{k}} \;\middle|\; \widehat{v}_{-\mathbf{k}} = \overline{\widehat{v}_{\mathbf{k}}},\ \forall \mathbf{k} \right\} \subset \bigcap_{s\in\mathbb{R}} H^s_\#(\Omega).$$

For all $s \in \mathbb{R}$, for all $r \leq s$, and for each $v \in H^s_\#(\Omega)$, the best approximation of v in $X_{N_{\mathrm{c}}}$ in any $H^r_\#$-norm is

$$\Pi_{N_{\mathrm{c}}} v = \sum_{\mathbf{k}\in\mathcal{R}^*, |\mathbf{k}| \leq \frac{2\pi}{L}N_{\mathrm{c}}} \widehat{v}_{\mathbf{k}} e_{\mathbf{k}}.$$

We denote by $\Pi^\perp_{N_{\mathrm{c}}} = (1 - \Pi_{N_{\mathrm{c}}})$ the orthogonal projector on $X^\perp_{N_{\mathrm{c}}}$, the orthogonal of $X_{N_{\mathrm{c}}}$. The variational approximation to the ground-state energy in $X_{N_{\mathrm{c}}}$ is defined as

$$\mathcal{E}_{0,N_{\mathrm{c}}} = \inf\left\{\mathcal{E}(\Psi_{N_{\mathrm{c}}})\ ,\ \Psi_{N_{\mathrm{c}}} \in \mathcal{M} \cap [X_{N_{\mathrm{c}}}]^N\right\}, \tag{17}$$

with $\mathcal{E}$ and $\mathcal{M}$ defined in (1) and (2). Let $\lambda_{1,N_{\mathrm{c}}} \leq \lambda_{2,N_{\mathrm{c}}} \leq \cdots \leq \lambda_{\dim(X_{N_{\mathrm{c}}}),N_{\mathrm{c}}}$ be the $\dim(X_{N_{\mathrm{c}}})$ eigenvalues (counting multiplicites) of the Hermitian linear operator $\mathcal{H}_{N_{\mathrm{c}},\mathrm{proj}} : X_{N_{\mathrm{c}}} \to X_{N_{\mathrm{c}}}$ defined as

$$\mathcal{H}_{N_{\mathrm{c}},\mathrm{proj}} = \Pi_{N_{\mathrm{c}}}\mathcal{H}\,\Pi_{N_{\mathrm{c}}} = -\frac{1}{2}\Pi_{N_{\mathrm{c}}}\Delta\Pi_{N_{\mathrm{c}}} + \Pi_{N_{\mathrm{c}}}\mathcal{V}\Pi_{N_{\mathrm{c}}}. \tag{18}$$

Let $(\phi_{1,N_c},\dots,\phi_{N,N_c})$ be an orthonormal family of eigenvectors of $\mathcal{H}_{N_c,\mathrm{proj}}$ associated with the eigenvalues $\lambda_{1,N_c}\le\cdots\le\lambda_{N,N_c}$:

$$\mathcal{H}_{N_c,\mathrm{proj}}\phi_{i,N_c}=\lambda_{i,N_c}\phi_{i,N_c},\qquad \int_\Omega \phi_{i,N_c}\phi_{j,N_c}=\delta_{ij},\qquad \forall 1\le i,j\le N,$$

and let $\Phi_{N_c}:=(\phi_{1,N_c},\dots,\phi_{N,N_c})^T$. Then Φ_{N_c} is a minimizer of (17). Denoting by

$$\boxed{\gamma_{0,N_c}=\sum_{i=1}^N |\phi_{i,N_c}\rangle\langle\phi_{i,N_c}|} \tag{19}$$

the associated density matrix, we have

$$\boxed{\mathcal{E}_{0,N_c}=\mathrm{Tr}\left(\mathcal{H}\,\gamma_{0,N_c}\right)=\sum_{j=1}^N \lambda_{j,N_c}.} \tag{20}$$

2.4 *A priori* results on the density matrices

From now on, we make the following technical assumption:

Assumption 2.2. *$\mathcal{V}$ is a $\mathcal{R}-$periodic potential satisfying $\mathcal{V}\in L^s_\#(\Omega)$ for $s>3/2$.*

Note that this assumption implies that $\mathcal{V}\in L^\infty_\#(\Omega)$ and $\nabla\mathcal{V}\in L^3_\#(\Omega)$.

The *a priori* error estimates established in [3] for the nonlinear Kohn–Sham model also hold true for the linear subproblem. In order to use these results in the present setting, it is convenient to reformulate them in terms of density matrices. As in [3], we introduce

$$\mathcal{M}^{\Phi^0}:=\left\{\Psi\in\mathcal{M}\;\middle|\;\|\Psi-\Phi^0\|_{L^2_\#}=\min_{U\in O(N)}\|U\Psi-\Phi^0\|_{L^2_\#}\right\},$$

where $\Phi^0=(\phi^0_1,\dots,\phi^0_N)^T$, $(\phi^0_1,\dots,\phi^0_N)$ being a family of orthonormal eigenvectors of $\mathcal{H}$ associated with the eigenvalues $\lambda^0_1\le\cdots\le\lambda^0_N$ fixed once and for all.

Proceeding as in [3], it can be shown that, for N_c large enough, (17) has a unique minimizer $\Phi^0_{N_c}=(\phi^0_{1,N_c},\dots,\phi^0_{N,N_c})^T$ belonging to $\mathcal{M}^{\Phi^0}$, that the set of minimizers of (17) is $O(N)\Phi^0_{N_c}$, and that, consequently, all the minimizers of (17) share the same density matrix. In particular

$$\gamma_{0,N_c}=\sum_{i=1}^N |\phi^0_{i,N_c}\rangle\langle\phi^0_{i,N_c}|.$$

We denote by

$$\Lambda^0_{N_c}=(\lambda^0_{ij,N_c})_{1\le i,j\le N}:=(\langle\phi^0_{i,N_c}|\mathcal{H}|\phi^0_{j,N_c}\rangle)_{1\le i,j\le N}\in\mathbb{R}^{N\times N} \tag{21}$$

the Lagrange multiplier matrix of the orthonormality constraints. Note that the matrix $\Lambda^0_{N_c}$ is not diagonal in general, but that we have

$$\mathcal{E}_{0,N_c}=\mathrm{Tr}\left(\Lambda^0_{N_c}\right).$$

The following lemma allows one to translate the *a priori* results of [3, Theorem 4.2] in terms of density matrices.

Lemma 2.3. *Under Assumption 2.1, there exist $0<c\le C<\infty$ and $N_c^0\in\mathbb{N}$, such that for all $N_c\ge N_c^0$,*

$$\|\Phi^0_{N_c}-\Phi^0\|_{L^2_\#}\;\le\;\|\gamma_{0,N_c}-\gamma_0\|_{\mathfrak{S}_2(L^2_\#)}\;\le\;\sqrt{2}\,\|\Phi^0_{N_c}-\Phi^0\|_{L^2_\#}, \tag{22}$$

$$c\|(1-\Delta)^{1/2}(\Phi^0_{N_c}-\Phi^0)\|_{L^2_\#}\;\le\;\|(1-\Delta)^{1/2}(\gamma_{0,N_c}-\gamma_0)\|_{\mathfrak{S}_2(L^2_\#)}\;\le\;C\|(1-\Delta)^{1/2}(\Phi^0_{N_c}-\Phi^0)\|_{L^2_\#}. \tag{23}$$

The proof is given in the Appendix.

We then immediately infer from [3, Theorem 4.2] that under Assumptions 2.1 and 2.2, there exist $C \in \mathbb{R}_+$ and $N_{\rm c}^0 \in \mathbb{N}$ such that for all $N_{\rm c} \geq N_{\rm c}^0$,

$$\|\gamma_0 - \gamma_{0,N_{\rm c}}\|_{\mathfrak{S}_2(L^2_\#)} \leq C N_{\rm c}^{-2}, \tag{24}$$

$$\|\gamma_0 - \gamma_{0,N_{\rm c}}\|_{\mathfrak{S}_2(L^2_\#)} \leq C N_{\rm c}^{-1} \|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_{\rm c}})\|_{\mathfrak{S}_2(L^2_\#)}. \tag{25}$$

Moreover, there exists $C \in \mathbb{R}_+$ such that

$$\|\Lambda^0 - \Lambda^0_{N_{\rm c}}\|_{\rm F} \leq C \|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_{\rm c}})\|^2_{\mathfrak{S}_2(L^2_\#)}, \tag{26}$$

where $\|\,.\,\|_{\rm F}$ denotes the Frobenius norm.

3 Post-processing of the plane-wave approximation

3.1 A key remark

Let us introduce the Hamiltonian on $L^2_\#(\Omega)$ with domain $H^2_\#(\Omega)$ defined by

$$\forall u \in H^2_\#(\Omega), \quad \mathcal{H}_{N_{\rm c}} u = -\frac{1}{2}\Delta u + \Pi_{N_{\rm c}} \mathcal{V} \Pi_{N_{\rm c}} u.$$

Since $X_{N_{\rm c}}$ and $X_{N_{\rm c}}^\perp$ are invariant subspaces of $\mathcal{H}_{N_{\rm c}}$, the Hamiltonian $\mathcal{H}_{N_{\rm c}}$ can be represented in term of $\mathcal{H}_{N_{\rm c},\rm proj}$ as follows:

$$\mathcal{H}_{N_{\rm c}} = \underbrace{\begin{pmatrix} \boxed{\mathcal{H}_{N_{\rm c},\rm proj}} & 0 \\ 0 & \boxed{-\tfrac{1}{2}\Delta} \end{pmatrix}}_{\begin{array}{cc} X_{N_{\rm c}} & X_{N_{\rm c}}^\perp \end{array}} \begin{array}{l} \Big\} X_{N_{\rm c}} \\ \Big\} X_{N_{\rm c}}^\perp \end{array} \tag{27}$$

The eigenvalues of the Laplace operator, which is diagonal in plane-wave bases, are explicitly known and its smallest eigenvalue on the invariant subspace $X_{N_{\rm c}}^\perp$ is larger than $\frac{1}{2}\left(\frac{LN_{\rm c}}{\pi}\right)^2$. Therefore, as soon as

$$\lambda_{N,N_{\rm c}} < \frac{1}{2}\left(\frac{LN_{\rm c}}{\pi}\right)^2, \tag{28}$$

where we recall that $\lambda_{1,N_{\rm c}} \leq \cdots \leq \lambda_{N,N_{\rm c}}$ are the lowest N eigenvalues (counting multiplicities) of the operator $\mathcal{H}_{N_{\rm c},\rm proj}$ defined in (18), we have

$$\forall j = 1,\ldots,N, \quad \mathcal{H}_{N_{\rm c}} \phi_{j,N_{\rm c}} = \lambda_{j,N_{\rm c}}\, \phi_{j,N_{\rm c}}, \qquad \text{and} \qquad \forall i,j = 1,\ldots,N, \quad \int_\Omega \phi_{i,N_{\rm c}}\, \phi_{j,N_{\rm c}} = \delta_{ij}, \tag{29}$$

and $\lambda_{1,N_{\rm c}} \leq \cdots \leq \lambda_{N,N_{\rm c}}$ are also the lowest N eigenvalues (counting multiplicities) of the operator $\mathcal{H}_{N_{\rm c}}$. A key observation is that the lowest energy eigenmodes of $\mathcal{H}$ satisfy

$$\forall j = 1,\ldots,N, \quad (\mathcal{H}_{N_{\rm c}} + \mathcal{V}_{N_{\rm c}}^\perp)\, \phi_j^0 = \lambda_j^0\, \phi_j^0, \qquad \text{and} \qquad \forall i,j = 1,\ldots,N, \quad \int_\Omega \phi_i^0\, \phi_j^0 = \delta_{ij}, \tag{30}$$

where

$$\mathcal{V}_{N_c}^{\perp} = \mathcal{V} - \Pi_{N_c} \mathcal{V} \Pi_{N_c}. \tag{31}$$

We can therefore apply the Rayleigh–Schrödinger perturbation method [13] using $(\phi_{j,N_c}, \lambda_{j,N_c})_{j=1,\dots,N}$ as the reference solution and $(\phi_j^0, \lambda_j^0)_{j=1,\dots,N}$ as the perturbed solution, in order to build improved approximations of the orbitals and eigenvalues respectively denoted by $(\widetilde{\phi_{j,N_c}})_{j=1,\dots,N}$ and $(\widetilde{\lambda_{j,N_c}})_{j=1,\dots,N}$, as well as an improved density matrix $\widetilde{\gamma_{N_c}}$ and improved energy $\widetilde{\mathcal{E}_{0,N_c}}$.

3.2 Corrections computation

More precisely, we apply first-order perturbation to the analytic family of operators $\mathcal{H}(\beta) = \mathcal{H}_{N_c} + \beta \mathcal{V}_{N_c}^{\perp}$, where $\beta \in \mathbb{R}$ is a parameter, which amounts to considering $\mathcal{H}(0) = \mathcal{H}_{N_c}$ and (29) as the unperturbed eigenvalue problem, and $\mathcal{H}(1) = \mathcal{H}$ and (30) as the perturbed eigenvalue problem. Assuming that the eigenvalues are not degenerate, we obtain at first order for the eigenfunctions, and at second order for the eigenvalues,

$$\forall\, j = 1, \dots, N, \qquad \phi_j^0 \simeq \phi_{j,N_c}^0 + \phi_{j,N_c}^{(1)}, \qquad \lambda_j^0 \simeq \lambda_{j,N_c} + \lambda_{j,N_c}^{(2)},$$

where

$$\phi_{j,N_c}^{(1)} = -\left(-\frac{1}{2}\Delta - \lambda_{j,N_c}\right)^{-1} r_j \in X_{N_c}^{\perp}, \tag{32}$$

with r_j being the residual

$$r_j = \left(-\frac{1}{2}\Delta + \mathcal{V} - \lambda_{j,N_c}\right)\phi_{j,N_c} = \left(\mathcal{H}_{N_c} + \mathcal{V}_{N_c}^{\perp} - \lambda_{j,N_c}\right)\phi_{j,N_c} = \mathcal{V}_{N_c}^{\perp}\phi_{j,N_c}, \tag{33}$$

and

$$\lambda_{j,N_c}^{(2)} = \langle \phi_{j,N_c}^{(1)} | r_j \rangle = -\langle r_j | \left(-\frac{1}{2}\Delta - \lambda_{j,N_c}\right)^{-1} | r_j \rangle. \tag{34}$$

We observe that the corrections on the eigenfunctions given in (32) are well-defined even if λ_{j,N_c} is degenerate. We therefore define the perturbed eigenvectors, density matrix, and energy for the general case as follows.

Definition 3.1 (Perturbed eigenvectors, eigenvalues, density matrix, and energy). *For all $N_c \geq N_c^0$ and all $j = 1, \dots, N$, the perturbed eigenvectors are defined as*

$$\widetilde{\phi_{j,N_c}} = \phi_{j,N_c} + \phi_{j,N_c}^{(1)},$$

the perturbed eigenvalues as

$$\widetilde{\lambda_{j,N_c}} = \lambda_{j,N_c} + \lambda_{j,N_c}^{(2)},$$

the perturbed density matrix as

$$\boxed{\widetilde{\gamma_{N_c}} = \gamma_{0,N_c} + \gamma_{N_c}^{(1)},}$$

with

$$\gamma_{N_c}^{(1)} = \sum_{j=1}^{N} |\phi_{j,N_c}^{(1)}\rangle\langle\phi_{j,N_c}| + \sum_{j=1}^{N} |\phi_{j,N_c}\rangle\langle\phi_{j,N_c}^{(1)}|, \tag{35}$$

and the perturbed energy as

$$\boxed{\widetilde{\mathcal{E}_{0,N_c}} = \sum_{j=1}^{N} \widetilde{\lambda_{j,N_c}} = \mathrm{Tr}\left(\gamma_{0,N_c}\mathcal{H}\,\widetilde{\gamma_{N_c}}\right).} \tag{36}$$

Remark 3.1. *Note that even if we call $\widetilde{\gamma_{N_c}}$ a density matrix, $\widetilde{\gamma_{N_c}} \notin \mathcal{K}$ in general. Indeed, $\widetilde{\gamma_{N_c}} = \widetilde{\gamma_{N_c}}^*$ and $\mathrm{Tr}\left(\widetilde{\gamma_{N_c}}\right) = N$, but we do not have in general $0 \leq \widetilde{\gamma_{N_c}} \leq 1$. Hence, the perturbed energy, which is defined as the sum of the perturbed eigenvalues, is not equal to the energy of the perturbed density matrix, i.e. $\widetilde{\mathcal{E}_{0,N_c}} \neq \mathrm{Tr}\left(\mathcal{H}\,\widetilde{\gamma_{N_c}}\right)$.*

Remark 3.2. *As we shall see in Section 5, the quantities $\phi_{j,N_c}^{(1)}$ are easily computable.*

4 Convergence improvement on the density matrix and the energy

4.1 Main results

The main results of this article are collected in the following theorem.

Theorem 4.1. *Under Assumptions 2.1–2.2, there exist $C \in \mathbb{R}_+$ and $N_{\rm c}^0 \in \mathbb{N}$, such that for all $N_{\rm c} \geq N_{\rm c}^0$,*

$$\boxed{\|(1-\Delta)^{1/2}(\widetilde{\gamma_{N_{\rm c}}} - \gamma_0)\|_{\mathfrak{S}_2(L^2_\#)} \leq C N_{\rm c}^{-2} \|(1-\Delta)^{1/2}(\gamma_{0,N_{\rm c}} - \gamma_0)\|_{\mathfrak{S}_2(L^2_\#)},} \tag{37}$$

and

$$\boxed{\left|\widetilde{\mathcal{E}_{0,N_{\rm c}}} - \mathcal{E}_0\right| \leq C N_{\rm c}^{-2} \left|\mathcal{E}_{0,N_{\rm c}} - \mathcal{E}_0\right|.} \tag{38}$$

4.2 Proofs

In order to prove Theorem 4.1, we first provide in Section 4.2.1 a decomposition of γ_0 based on spectral projection in Lemma 4.3, relying on Lemma 4.2 for a rigorous justification of the contour integral. In Section 4.2.2, we decompose the difference $\gamma_0 - \widetilde{\gamma_{N_{\rm c}}}$ into three parts in Lemma 4.4, and we then estimate each of these terms in three of the following Lemmas 4.5, 4.7, and 4.8, relying on an intermediary estimate presented in Lemma 4.6, in order to prove estimate (37). Finally, in Section 4.2.3, we provide a proof for estimate (38).

4.2.1 Exact density matrix in terms of approximate density matrix

Lemma 4.2. *Let Γ be the circle in the complex plane symmetric with respect to the real axis and containing the real numbers $\lambda_1^0 - 1$ and $\epsilon_{\rm F}$. There exists $N_{\rm c}^0 \in \mathbb{N}$ such that for all $N_{\rm c} \geq N_{\rm c}^0$, Γ encloses the lowest N eigenvalues of both the operators $\mathcal{H}$ and $\mathcal{H}_{N_{\rm c}}$, and none of the higher ones.*

Proof. As $\mathcal{H}$ and $\mathcal{H}_{N_{\rm c}}$ are self-adjoint operators, their eigenvalues noted respectively $(\lambda_i^0)_{i\in\mathbb{N}^*}$ and $(\lambda_{i,N_{\rm c}})_{i\in\mathbb{N}^*}$ (with increasing values and counting multiplicities) are real. From the gap assumption 2.1, and the definition of the Fermi level, we have

$$\forall i = 1, \ldots, N, \quad \lambda_i^0 < \epsilon_F, \quad \text{and} \quad \forall i > N, \quad \lambda_i^0 > \epsilon_F. \tag{39}$$

The plane-wave discretization being variational, there holds

$$\forall i = 1, \ldots, \dim(X_{N_{\rm c}}), \quad \lambda_i^0 \leq \lambda_{i,N_{\rm c}}.$$

Moreover, classical convergence results (see e.g. [8, Chapter 5]) guarantee that

$$\max_{i=1,\ldots,N} |\lambda_{i,N_{\rm c}} - \lambda_i^0| \underset{N_{\rm c}\to+\infty}{\longrightarrow} 0.$$

Therefore, there exists $N_{\rm c}^0 \in \mathbb{N}$ such that for all $N_{\rm c} \geq N_{\rm c}^0$, $\lambda_{N,N_{\rm c}} \leq \lambda_{N,N_{\rm c}^0} < \epsilon_F$, and the eigenvalues of the Laplace operator on $X_{N_{\rm c}}^\perp$ are larger than $\lambda_{N+1}^0 > \epsilon_F$, so that

$$\forall N_{\rm c} \geq N_{\rm c}^0, \qquad \forall i = 1, \ldots, N, \quad \lambda_{i,N_{\rm c}} \leq \lambda_{N,N_{\rm c}^0} < \epsilon_F, \quad \text{and} \quad \forall i > N, \quad \lambda_{i,N_{\rm c}} \geq \lambda_{N+1}^0 > \epsilon_F. \tag{40}$$

Combining (39) and (40) concludes the proof of the lemma. □

Using the Cauchy residue theorem and functional calculus for self-adjoint operators, the ground-state density matrix of $\mathcal{H}$ can be written as

$$\gamma_0 = \frac{1}{2\pi i} \oint_\Gamma (z - \mathcal{H})^{-1} \, dz = \frac{1}{2\pi i} \oint_\Gamma \left(z - \mathcal{H}_{N_{\rm c}} - \mathcal{V}_{N_{\rm c}}^\perp\right)^{-1} dz. \tag{41}$$

Since $\mathcal{V} \in L^\infty_\#(\Omega)$, $\mathcal{V}^\perp_{N_\mathrm{c}}$ is $\mathcal{H}_{N_\mathrm{c}}$-bounded, and we can perform a Dyson expansion of (41) at second order. We obtain

$$\begin{aligned}\gamma_0 = \frac{1}{2\pi i}\oint_\Gamma (z-\mathcal{H}_{N_\mathrm{c}})^{-1}dz + \frac{1}{2\pi i}\oint_\Gamma (z-\mathcal{H}_{N_\mathrm{c}})^{-1}\mathcal{V}^\perp_{N_\mathrm{c}}(z-\mathcal{H}_{N_\mathrm{c}})^{-1}dz \\ + \frac{1}{2\pi i}\oint_\Gamma (z-\mathcal{H})^{-1}\mathcal{V}^\perp_{N_\mathrm{c}}(z-\mathcal{H}_{N_\mathrm{c}})^{-1}\mathcal{V}^\perp_{N_\mathrm{c}}(z-\mathcal{H}_{N_\mathrm{c}})^{-1}dz, \end{aligned} \tag{42}$$

where each term of the right-hand side is well-defined.

Lemma 4.3 (Second order expansion of γ_0)**.** *There holds*

$$\gamma_0 = \gamma_{0,N_\mathrm{c}} + \gamma^{(1)}_{N_\mathrm{c}} + \widetilde{Q_{N_\mathrm{c}}}, \tag{43}$$

where $\gamma^{(1)}_{N_\mathrm{c}}$ is the finite-rank operator defined in (35) *and where*

$$\widetilde{Q_{N_\mathrm{c}}} := \frac{1}{2i\pi}\oint_\Gamma (z-\mathcal{H})^{-1}\mathcal{V}^\perp_{N_\mathrm{c}}(z-\mathcal{H}_{N_\mathrm{c}})^{-1}\mathcal{V}^\perp_{N_\mathrm{c}}(z-\mathcal{H}_{N_\mathrm{c}})^{-1}dz. \tag{44}$$

Proof. The operator $\mathcal{H}_{N_\mathrm{c}}$ being self-adjoint with compact resolvent, it can be diagonalized in an orthonormal basis. Hence, there exists a sequence $(\psi_k, \varepsilon_k)_{k\geq 1}$ with $(\psi_k)_{k\geq 1}$ an orthonormal basis of $L^2_\#(\Omega)$ consisting of functions of $H^2_\#(\Omega)$ and $(\varepsilon_k)_{k\geq 1}$ a non-decreasing sequence of real numbers such that

$$\forall k \geq 1, \quad \mathcal{H}_{N_\mathrm{c}}\psi_k = \varepsilon_k \psi_k.$$

Without loss of generality, we can choose a basis such that, in addition, for $k = 1, \dots, N$, $\psi_k = \phi_{k,N_\mathrm{c}}$ and $\varepsilon_k = \lambda_{k,N_\mathrm{c}}$. The operator $\mathcal{H}_{N_\mathrm{c}}$ can then be written as

$$\mathcal{H}_{N_\mathrm{c}} = \sum_{k\geq 1} \varepsilon_k |\psi_k\rangle\langle\psi_k|.$$

Let us show that the expansions (42) and (43) are identical. First, we have

$$\gamma_{0,N_\mathrm{c}} = \frac{1}{2\pi i}\oint_\Gamma (z-\mathcal{H}_{N_\mathrm{c}})^{-1}\, dz.$$

Let us now prove that the second term in the right hand side, that is

$$\gamma^{(1)} := \frac{1}{2\pi i}\oint_\Gamma (z-\mathcal{H}_{N_\mathrm{c}})^{-1}\,\mathcal{V}^\perp_{N_\mathrm{c}}\,(z-\mathcal{H}_{N_\mathrm{c}})^{-1}\, dz, \tag{45}$$

is in fact equal to the operator $\gamma^{(1)}_{N_\mathrm{c}}$ defined in (35). We have

$$\forall k,l \in \mathbb{N}^*, \quad \langle\psi_k|\gamma^{(1)}_{N_\mathrm{c}}|\psi_l\rangle = \sum_{j=1}^N \langle\psi_k|\phi^{(1)}_{j,N_\mathrm{c}}\rangle\langle\phi_{j,N_\mathrm{c}}|\psi_l\rangle + \sum_{j=1}^N \langle\psi_k|\phi_{j,N_\mathrm{c}}\rangle\langle\phi^{(1)}_{j,N_\mathrm{c}}|\psi_l\rangle.$$

As for all $j = 1, \dots, N$, $\phi_{j,N_\mathrm{c}} \in X_{N_\mathrm{c}}$ and $\phi^{(1)}_{j,N_\mathrm{c}} \in X^\perp_{N_\mathrm{c}}$, we have $\langle\psi_k|\gamma^{(1)}_{N_\mathrm{c}}|\psi_l\rangle = 0$ for all $k,l \in \mathbb{N}^*$ such that either $k,l > N$, or $k,l \leq N$. Moreover, for all $k \leq N$ and $l > N$, we have

$$\begin{aligned}\langle\psi_k|\gamma^{(1)}_{N_\mathrm{c}}|\psi_l\rangle = \langle\phi_{k,N_\mathrm{c}}|\gamma^{(1)}_{N_\mathrm{c}}|\psi_l\rangle = \langle\phi^{(1)}_{k,N_\mathrm{c}}|\psi_l\rangle = -\langle\phi_{k,N_\mathrm{c}}|\mathcal{V}^\perp_{N_\mathrm{c}}(\mathcal{H}_{N_\mathrm{c}}-\lambda_{k,N_\mathrm{c}})^{-1}|\psi_l\rangle \\ = \frac{1}{\lambda_{k,N_\mathrm{c}}-\varepsilon_l}\langle\phi_{k,N_\mathrm{c}}|\mathcal{V}^\perp_{N_\mathrm{c}}|\psi_l\rangle = \frac{1}{\varepsilon_k-\varepsilon_l}\langle\psi_k|\mathcal{V}^\perp_{N_\mathrm{c}}|\psi_l\rangle,\end{aligned}$$

and likewise

$$\langle\psi_l|\gamma^{(1)}_{N_c}|\psi_k\rangle = \frac{1}{\varepsilon_l-\varepsilon_k}\langle\psi_l|\mathcal{V}^{\perp}_{N_c}|\psi_k\rangle.$$

Thus, for all $k,l\in\mathbb{N}^*$, we have

$$\langle\psi_k|\gamma^{(1)}_{N_c}|\psi_l\rangle = \frac{1}{\varepsilon_k-\varepsilon_l}\left(\mathbb{1}_{k\leq N}\mathbb{1}_{l>N}-\mathbb{1}_{k>N}\mathbb{1}_{l\leq N}\right)\langle\psi_k|\mathcal{V}^{\perp}_{N_c}|\psi_l\rangle.$$

On the other hand, for all $k,l\in\mathbb{N}^*$, using the Cauchy residue theorem,

$$\begin{aligned}\langle\psi_k|\gamma^{(1)}|\psi_l\rangle &= \frac{1}{2\pi i}\oint_\Gamma \langle\psi_k|(z-\mathcal{H}_{N_c})^{-1}\mathcal{V}^{\perp}_{N_c}(z-\mathcal{H}_{N_c})^{-1}|\psi_l\rangle dz\\ &= \frac{1}{2\pi i}\left(\oint_\Gamma (z-\varepsilon_k)^{-1}(z-\varepsilon_l)^{-1}dz\right)\langle\psi_k|\mathcal{V}^{\perp}_{N_c}|\psi_l\rangle\\ &= \frac{1}{\varepsilon_k-\varepsilon_l}\left(\mathbb{1}_{k\leq N}\mathbb{1}_{l>N}-\mathbb{1}_{k>N}\mathbb{1}_{l\leq N}\right)\langle\psi_k|\mathcal{V}^{\perp}_{N_c}|\psi_l\rangle.\end{aligned}$$

Therefore, for all $k,l\in\mathbb{N}^*$, $\langle\psi_k|\gamma^{(1)}|\psi_l\rangle = \langle\psi_k|\gamma^{(1)}_{N_c}|\psi_l\rangle$. Finally $\gamma^{(1)}=\gamma^{(1)}_{N_c}$, and the definition of $\widetilde{Q_{N_c}}$ in (44) allows one to conclude the proof of the lemma. □

4.2.2 Proof of estimate (37)

Lemma 4.4. *There holds*

$$\gamma_0-\widetilde{\gamma_{N_c}} = (\gamma_0-\gamma_{0,N_c})^2+\widetilde{Q_{N_c}}\gamma_{0,N_c}+\gamma_{0,N_c}\widetilde{Q_{N_c}}, \tag{46}$$

with $\widetilde{Q_{N_c}}$ defined in (44).

Proof. Let us first remark, from (43), and the property $\gamma^2_{0,N_c}=\gamma_{0,N_c}$, that

$$\begin{aligned}\gamma_{0,N_c}\gamma_0 &= \gamma_{0,N_c}+\gamma_{0,N_c}\gamma^{(1)}_{N_c}+\gamma_{0,N_c}\widetilde{Q_{N_c}},\\ \gamma_0\gamma_{0,N_c} &= \gamma_{0,N_c}+\gamma^{(1)}_{N_c}\gamma_{0,N_c}+\widetilde{Q_{N_c}}\gamma_{0,N_c}.\end{aligned}$$

Moreover, as for all $i,j=1,2,\dots,N$, $\phi_{i,N_c}\in X_{N_c}$ and $\phi^{(1)}_{j,N_c}\in X^{\perp}_{N_c}$, ϕ_{i,N_c} is orthogonal to $\phi^{(1)}_{j,N_c}$, and therefore

$$\begin{aligned}\gamma_{0,N_c}\gamma^{(1)}_{N_c}+\gamma^{(1)}_{N_c}\gamma_{0,N_c} =&\gamma_{0,N_c}\left(\sum_{i=1}^N |\phi^{(1)}_{i,N_c}\rangle\langle\phi_{i,N_c}|+|\phi_{i,N_c}\rangle\langle\phi^{(1)}_{i,N_c}|\right)\\ &+\left(\sum_{i=1}^N |\phi^{(1)}_{i,N_c}\rangle\langle\phi_{i,N_c}|+|\phi_{i,N_c}\rangle\langle\phi^{(1)}_{i,N_c}|\right)\gamma_{0,N_c}\\ =&\gamma_{0,N_c}\sum_{i=1}^N |\phi_{i,N_c}\rangle\langle\phi^{(1)}_{i,N_c}|+\sum_{i=1}^N |\phi^{(1)}_{i,N_c}\rangle\langle\phi_{i,N_c}|\gamma_{0,N_c}\\ =&\sum_{i=1}^N |\phi_{i,N_c}\rangle\langle\phi^{(1)}_{i,N_c}|+\sum_{i=1}^N |\phi^{(1)}_{i,N_c}\rangle\langle\phi_{i,N_c}| = \gamma^{(1)}_{N_c}.\end{aligned}$$

Hence

$$\gamma_{0,N_c}\gamma_0+\gamma_0\gamma_{0,N_c} = 2\gamma_{0,N_c}+\gamma^{(1)}_{N_c}+\widetilde{Q_{N_c}}\gamma_{0,N_c}+\gamma_{0,N_c}\widetilde{Q_{N_c}} = \gamma_{0,N_c}+\widetilde{\gamma_{N_c}}+\widetilde{Q_{N_c}}\gamma_{0,N_c}+\gamma_{0,N_c}\widetilde{Q_{N_c}},$$

so that

$$(\gamma_0-\gamma_{0,N_c})^2 = \gamma_0-(\gamma_{0,N_c}\gamma_0+\gamma_0\gamma_{0,N_c})+\gamma_{0,N_c} = \gamma_0-\widetilde{\gamma_{N_c}}-\widetilde{Q_{N_c}}\gamma_{0,N_c}-\gamma_{0,N_c}\widetilde{Q_{N_c}},$$

from which we deduce (46). □

Lemma 4.5. *There exist $C \in \mathbb{R}_+$ and $N_{\rm c}^0 \in \mathbb{N}$, such that for all $N_{\rm c} \geq N_{\rm c}^0$,*

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

Proof. By cyclicity of the trace, noting that $(\gamma_0 - \gamma_{0,N_{\rm c}})$ is of finite rank, and using (11) and (10),

$$\begin{aligned}
\|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_{\rm c}})^2\|^2_{\mathfrak{S}_2(L^2_\#)} &= \mathrm{Tr}\,\left((\gamma_0 - \gamma_{0,N_{\rm c}})^2(1-\Delta)(\gamma_0 - \gamma_{0,N_{\rm c}})^2\right) \\
&= \mathrm{Tr}\,\left((\gamma_0 - \gamma_{0,N_{\rm c}})^2(\gamma_0 - \gamma_{0,N_{\rm c}})(1-\Delta)(\gamma_0 - \gamma_{0,N_{\rm c}})\right) \\
&\leq \|(\gamma_0 - \gamma_{0,N_{\rm c}})^2(\gamma_0 - \gamma_{0,N_{\rm c}})(1-\Delta)(\gamma_0 - \gamma_{0,N_{\rm c}})\|_{\mathfrak{S}_1(L^2_\#)} \\
&\leq \|\gamma_0 - \gamma_{0,N_{\rm c}}\|^2 \|(\gamma_0 - \gamma_{0,N_{\rm c}})(1-\Delta)(\gamma_0 - \gamma_{0,N_{\rm c}})\|_{\mathfrak{S}_1(L^2_\#)} \\
&\leq \|\gamma_0 - \gamma_{0,N_{\rm c}}\|^2_{\mathfrak{S}_2(L^2_\#)} \|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_{\rm c}})\|^2_{\mathfrak{S}_2(L^2_\#)}.
\end{aligned}$$

Using the *a priori* estimate of $\|\gamma_0 - \gamma_{0,N_{\rm c}}\|_{\mathfrak{S}_2(L^2_\#)}$ given in (24) finishes the proof. □

We now provide an estimate for $\|(1-\Delta)^{-1/2}\mathcal{V}^\perp_{N_{\rm c}}\gamma_{0,N_{\rm c}}\|_{\mathfrak{S}_2(L^2_\#)}$ which will be useful in the proof of Lemma 4.7 and estimate (38).

Lemma 4.6. *There exist $C \in \mathbb{R}_+$ and $N_{\rm c}^0 \in \mathbb{N}$, such that for $N_{\rm c} \geq N_{\rm c}^0$,*

$$\|(1-\Delta)^{-1/2}\mathcal{V}^\perp_{N_{\rm c}}\gamma_{0,N_{\rm c}}\|_{\mathfrak{S}_2(L^2_\#)} \leq C\|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_{\rm c}})\|_{\mathfrak{S}_2(L^2_\#)}. \tag{47}$$

Proof. Decomposing $\mathcal{V}^\perp_{N_{\rm c}}\gamma_{0,N_{\rm c}}$ as

$$\mathcal{V}^\perp_{N_{\rm c}}\gamma_{0,N_{\rm c}} \;=\; \mathcal{H}\,(\gamma_{0,N_{\rm c}} - \gamma_0) + \mathcal{H}\,\gamma_0 - \mathcal{H}_{N_{\rm c}}\gamma_{0,N_{\rm c}},$$

we get

$$\begin{aligned}
\|(1-\Delta)^{-1/2}\mathcal{V}^\perp_{N_{\rm c}}\gamma_{0,N_{\rm c}}\|_{\mathfrak{S}_2(L^2_\#)} \leq{}& \|(1-\Delta)^{-1/2}\mathcal{H}\,(\gamma_{0,N_{\rm c}} - \gamma_0)\|_{\mathfrak{S}_2(L^2_\#)} \\
&+ \|(1-\Delta)^{-1/2}(\mathcal{H}\,\gamma_0 - \mathcal{H}_{N_{\rm c}}\gamma_{0,N_{\rm c}})\|_{\mathfrak{S}_2(L^2_\#)}.
\end{aligned} \tag{48}$$

Since $(1-\Delta)^{-1/2}\mathcal{H}\,(1-\Delta)^{-1/2}$ is a bounded operator (see e.g. [5, Lemma 1] for a proof of this classical result) and from (12), there exists $C \in \mathbb{R}_+$ such that for all $N_{\rm c} \in \mathbb{N}$,

$$\|(1-\Delta)^{-1/2}\mathcal{H}\,(\gamma_{0,N_{\rm c}} - \gamma_0)\|_{\mathfrak{S}_2(L^2_\#)} \;\leq\; C\|(1-\Delta)^{1/2}(\gamma_{0,N_{\rm c}} - \gamma_0)\|_{\mathfrak{S}_2(L^2_\#)}. \tag{49}$$

In order to bound the second term of the right-hand side of (48), we first rewrite the operator $\mathcal{H}\,\gamma_0 - \mathcal{H}_{N_{\rm c}}\gamma_{0,N_{\rm c}}$ as follows, denoting by $\lambda^0_{ij} = \lambda^0_i\delta_{ij}$, and using $\lambda^0_{ij,N_{\rm c}}$ defined in (21):

$$\begin{aligned}
\mathcal{H}\,\gamma_0 - \mathcal{H}_{N_{\rm c}}\gamma_{0,N_{\rm c}} &= \sum_{i=1}^N \lambda^0_i |\phi^0_i\rangle\langle\phi^0_i| - \sum_{i,j=1}^N \lambda^0_{ij,N_{\rm c}} |\phi^0_{i,N_{\rm c}}\rangle\langle\phi^0_{j,N_{\rm c}}| \\
&= \sum_{i=1}^N \lambda^0_i \left(|\phi^0_i\rangle\langle\phi^0_i| - |\phi^0_{i,N_{\rm c}}\rangle\langle\phi^0_{i,N_{\rm c}}|\right) + \sum_{i=1}^N \lambda^0_i |\phi^0_{i,N_{\rm c}}\rangle\langle\phi^0_{i,N_{\rm c}}| - \sum_{i,j=1}^N \lambda^0_{ij,N_{\rm c}} |\phi^0_{i,N_{\rm c}}\rangle\langle\phi^0_{j,N_{\rm c}}| \\
&= \sum_{i=1}^N \lambda^0_i \left(|\phi^0_i\rangle\langle\phi^0_i| - |\phi^0_{i,N_{\rm c}}\rangle\langle\phi^0_{i,N_{\rm c}}|\right) + \sum_{i,j=1}^N (\lambda^0_{ij} - \lambda^0_{ij,N_{\rm c}}) |\phi^0_{i,N_{\rm c}}\rangle\langle\phi^0_{j,N_{\rm c}}|.
\end{aligned}$$

Using the triangle and the Cauchy–Schwarz inequality, we get

$$\begin{aligned}
\|\mathcal{H}\,\gamma_0 - \mathcal{H}_{N_\mathrm{c}}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)} &\leq \left\|\sum_{i=1}^N \lambda_i^0 \left(|\phi_i^0\rangle\langle\phi_i^0| - |\phi_{i,N_\mathrm{c}}^0\rangle\langle\phi_{i,N_\mathrm{c}}^0|\right)\right\|_{\mathfrak{S}_2(L^2_\#)} \\
&\quad + \left\|\sum_{i,j=1}^N (\lambda_{ij}^0 - \lambda_{ij,N_\mathrm{c}}^0)|\phi_{i,N_\mathrm{c}}^0\rangle\langle\phi_{j,N_\mathrm{c}}^0|\right\|_{\mathfrak{S}_2(L^2_\#)} \\
&\leq \sum_{i=1}^N |\lambda_i^0| \left\|\left(|\phi_i^0\rangle\langle\phi_i^0| - |\phi_{i,N_\mathrm{c}}^0\rangle\langle\phi_{i,N_\mathrm{c}}^0|\right)\right\|_{\mathfrak{S}_2(L^2_\#)} \\
&\quad + \sum_{i,j=1}^N |\lambda_{ij}^0 - \lambda_{ij,N_\mathrm{c}}^0| \left\||\phi_{i,N_\mathrm{c}}^0\rangle\langle\phi_{j,N_\mathrm{c}}^0|\right\|_{\mathfrak{S}_2(L^2_\#)} \\
&\leq \left(\sum_{i=1}^N |\lambda_i^0|^2\right)^{1/2} \left(\sum_{i=1}^N \left\||\phi_i^0\rangle\langle\phi_i^0| - |\phi_{i,N_\mathrm{c}}^0\rangle\langle\phi_{i,N_\mathrm{c}}^0|\right\|^2_{\mathfrak{S}_2(L^2_\#)}\right)^{1/2} \\
&\quad + \|\Lambda^0 - \Lambda^0_{N_\mathrm{c}}\|_F \left(\sum_{i,j=1}^N \left\||\phi_{i,N_\mathrm{c}}^0\rangle\langle\phi_{j,N_\mathrm{c}}^0|\right\|^2_{\mathfrak{S}_2(L^2_\#)}\right)^{1/2} \\
&\leq 2\left(\sum_{i=1}^N |\lambda_i^0|^2\right)^{1/2} \|\Phi^0 - \Phi^0_{N_\mathrm{c}}\|_{L^2_\#} + N\|\Lambda^0 - \Lambda^0_{N_\mathrm{c}}\|_F.
\end{aligned}$$

Using (22), (25), and (26), we obtain that there exist $C \in \mathbb{R}_+$ and $N_\mathrm{c}^0 \in \mathbb{N}$ such that for all $N_\mathrm{c} \geq N_\mathrm{c}^0$,

$$\|\mathcal{H}\,\gamma_0 - \mathcal{H}_{N_\mathrm{c}}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)} \leq CN_\mathrm{c}^{-1}\|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_\mathrm{c}})\|_{\mathfrak{S}_2(L^2_\#)} + C\|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_\mathrm{c}})\|^2_{\mathfrak{S}_2(L^2_\#)}.$$

Since $\|(1-\Delta)^{-1/2}\| \leq 1$, this shows in particular that

$$\|(1-\Delta)^{-1/2}(\mathcal{H}\,\gamma_0 - \mathcal{H}_{N_\mathrm{c}}\gamma_{0,N_\mathrm{c}})\|_{\mathfrak{S}_2(L^2_\#)} \leq C\|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_\mathrm{c}})\|_{\mathfrak{S}_2(L^2_\#)}. \tag{50}$$

Inserting (49) and (50) in (48) concludes the proof of the lemma. □

Lemma 4.7. *There exist $C \in \mathbb{R}_+$ and $N_\mathrm{c}^0 \in \mathbb{N}$ such that for all $N_\mathrm{c} \geq N_\mathrm{c}^0$,*

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

Proof. Using the definition (44) and the fact that $(z - \mathcal{H}_{N_\mathrm{c}})^{-1}$ and γ_{0,N_c} commute, we obtain

$$(1-\Delta)^{1/2}\widetilde{Q_{N_\mathrm{c}}}\gamma_{0,N_\mathrm{c}} = \frac{1}{2i\pi}\oint_\Gamma (1-\Delta)^{1/2}(z-\mathcal{H})^{-1}\mathcal{V}_{N_\mathrm{c}}^\perp(z-\mathcal{H}_{N_\mathrm{c}})^{-1}\mathcal{V}_{N_\mathrm{c}}^\perp\gamma_{0,N_\mathrm{c}}(z-\mathcal{H}_{N_\mathrm{c}})^{-1}dz.$$

Since $\mathrm{Ran}(\mathcal{V}_{N_\mathrm{c}}^\perp\gamma_{0,N_\mathrm{c}}) \subset X_{N_\mathrm{c}}^\perp$, we have $\mathcal{V}_{N_\mathrm{c}}^\perp\gamma_{0,N_\mathrm{c}} = \Pi_{N_\mathrm{c}}^\perp\mathcal{V}_{N_\mathrm{c}}^\perp\gamma_{0,N_\mathrm{c}}$. Observing that

$$(z-\mathcal{H}_{N_\mathrm{c}})^{-1}\Pi_{N_\mathrm{c}}^\perp = \Pi_{N_\mathrm{c}}^\perp\left(z + \frac{1}{2}\Delta_{|X_{N_\mathrm{c}}^\perp}\right)^{-1}\Pi_{N_\mathrm{c}}^\perp,$$

we thus obtain

$$(z-\mathcal{H}_{N_\mathrm{c}})^{-1}\mathcal{V}_{N_\mathrm{c}}^\perp\gamma_{0,N_\mathrm{c}} = (z-\mathcal{H}_{N_\mathrm{c}})^{-1}\Pi_{N_\mathrm{c}}^\perp\mathcal{V}_{N_\mathrm{c}}^\perp\gamma_{0,N_\mathrm{c}} = \Pi_{N_\mathrm{c}}^\perp\left(z + \frac{1}{2}\Delta_{|X_{N_\mathrm{c}}^\perp}\right)^{-1}\mathcal{V}_{N_\mathrm{c}}^\perp\gamma_{0,N_\mathrm{c}}.$$

Therefore,

$$(1-\Delta)^{1/2}\widetilde{Q_{N_\mathrm{c}}}\gamma_{0,N_\mathrm{c}} = \frac{1}{2i\pi}\oint_\Gamma \left[(1-\Delta)^{1/2}(z-\mathcal{H})^{-1}(1-\Delta)^{1/2}\right]\ \left[(1-\Delta)^{-1/2}\mathcal{V}^\perp_{N_\mathrm{c}}(1-\Delta)^{-1/2}\Pi^\perp_{N_\mathrm{c}}\right] \times \left[\Pi^\perp_{N_\mathrm{c}}(1-\Delta)^{1/2}(z+\frac{1}{2}\Delta)^{-1}(1-\Delta)^{1/2}\Pi^\perp_{N_\mathrm{c}}\right] \times \left[(1-\Delta)^{-1/2}\mathcal{V}^\perp_{N_\mathrm{c}}\gamma_{0,N_\mathrm{c}}(z-\mathcal{H}_{N_\mathrm{c}})^{-1}\right] dz. \tag{51}$$

First,

$$\begin{aligned}\|(1-\Delta)^{-1/2}\mathcal{V}^\perp_{N_\mathrm{c}}(1-\Delta)^{-1/2}\Pi^\perp_{N_\mathrm{c}}\| &= \|(1-\Delta)^{-1/2}\mathcal{V}(1-\Delta)^{-1/2}\Pi^\perp_{N_\mathrm{c}}\| \\ &= \|\Pi^\perp_{N_\mathrm{c}}(1-\Delta)^{-1/2}\mathcal{V}(1-\Delta)^{-1/2}\| \\ &\leq \|\Pi^\perp_{N_\mathrm{c}}(1-\Delta)^{-1}\|\|(1-\Delta)^{1/2}\mathcal{V}(1-\Delta)^{-1/2}\|.\end{aligned}$$

Since $\|(1-\Delta)^{1/2}\mathcal{V}(1-\Delta)^{-1/2}\|$ equals the operator norm of $\mathcal{V}$, considered as a multiplicative operator from $H^1_\#(\Omega)$ to $H^1_\#(\Omega)$, it can be shown using classical Sobolev embeddings that there exists $C\in\mathbb{R}_+$ such that

$$\|(1-\Delta)^{1/2}\mathcal{V}(1-\Delta)^{-1/2}\| \leq C\left(\|\mathcal{V}\|_{L^\infty}+\|\nabla\mathcal{V}\|_{L^3}\right),$$

which is bounded under Assumption 2.2. Since $\|\Pi^\perp_{N_\mathrm{c}}(1-\Delta)^{-1}\| \leq (1+N_\mathrm{c}^2)^{-1}$ for all $N_\mathrm{c}\in\mathbb{N}$, there exist $C\in\mathbb{R}_+$ and $N_\mathrm{c}^0\in\mathbb{N}$ such that for all $N_\mathrm{c}\geq N_\mathrm{c}^0$,

$$\|(1-\Delta)^{-1/2}\mathcal{V}^\perp_{N_\mathrm{c}}(1-\Delta)^{-1/2}\Pi^\perp_{N_\mathrm{c}}\| \leq CN_\mathrm{c}^{-2}. \tag{52}$$

Finally,

$$\begin{aligned}\max_{z\in\Gamma}\|(1-\Delta)^{-1/2}\mathcal{V}^\perp_{N_\mathrm{c}}\gamma_{0,N_\mathrm{c}}(z-\mathcal{H}_{N_\mathrm{c}})^{-1}\|_{\mathfrak{S}_2(L^2_\#)} &= \left(\sum_{i=1}^N \max_{z\in\Gamma}|z-\lambda_{i,N_\mathrm{c}}|^{-2}\|(1-\Delta)^{-1/2}\mathcal{V}^\perp_{N_\mathrm{c}}\phi_{i,N_\mathrm{c}}\|^2_{L^2_\#}\right)^{1/2} \\ &\leq \max_{\substack{z\in\Gamma,\\ i=1,\dots,N}}|z-\lambda_{i,N_\mathrm{c}}|^{-2}\left(\sum_{i=1}^N\|(1-\Delta)^{-1/2}\mathcal{V}^\perp_{N_\mathrm{c}}\phi_{i,N_\mathrm{c}}\|^2_{L^2_\#}\right)^{1/2} \\ &= \max_{\substack{z\in\Gamma,\\ i=1,\dots,N}}|z-\lambda_{i,N_\mathrm{c}}|^{-2}\|(1-\Delta)^{-1/2}\mathcal{V}^\perp_{N_\mathrm{c}}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)}.\end{aligned}$$

From the definition of the contour Γ, $\max\limits_{\substack{z\in\Gamma,\\ i=1,\dots,N}}|z-\lambda_{i,N_\mathrm{c}}|^{-2}$ is bounded uniformly in N_c for N_c large enough. Hence, combining the above inequality with (47), we obtain that there exist $C\in\mathbb{R}_+$ and $N_\mathrm{c}^0\in\mathbb{N}$ such that for all $N_\mathrm{c}\geq N_\mathrm{c}^0$,

$$\max_{z\in\Gamma}\|(1-\Delta)^{-1/2}\mathcal{V}^\perp_{N_\mathrm{c}}\gamma_{0,N_\mathrm{c}}(z-\mathcal{H}_{N_\mathrm{c}})^{-1}\|_{\mathfrak{S}_2(L^2_\#)} \leq\ C\|(1-\Delta)^{1/2}(\gamma_{0,N_\mathrm{c}}-\gamma_0)\|_{\mathfrak{S}_2(L^2_\#)}. \tag{53}$$

We are now in position to estimate $\|(1-\Delta)^{1/2}\widetilde{Q_{N_\mathrm{c}}}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)}$. We start from (51). It is classical that $\max\limits_{z\in\Gamma}\|(1-\Delta)^{1/2}(z-\mathcal{H})^{-1}(1-\Delta)^{1/2}\|$ is bounded (see e.g. [5, Lemma 1]). Moreover, $\max\limits_{z\in\Gamma}\|\Pi^\perp_{N_\mathrm{c}}(1-\Delta)^{1/2}(z+\frac{1}{2}\Delta)^{-1}(1-\Delta)^{1/2}\Pi^\perp_{N_\mathrm{c}}\|$ is also bounded. Using estimates (52) and (53) allows one to conclude the proof of the lemma. □

Lemma 4.8. *There exist $C\in\mathbb{R}_+$ and $N_\mathrm{c}^0\in\mathbb{N}$ such that for all $N_\mathrm{c}\geq N_\mathrm{c}^0$,*

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

Proof. Noting that $\gamma_{0,N_\mathrm{c}}^2 = \gamma_{0,N_\mathrm{c}}$, using (13) and the cyclicity of the trace, we obtain

$$\begin{aligned}\|(1-\Delta)^{1/2}\gamma_{0,N_\mathrm{c}}\widetilde{Q_{N_\mathrm{c}}}\|_{\mathfrak{S}_2(L^2_\#)} &\leq \|(1-\Delta)^{1/2}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)}\|\gamma_{0,N_\mathrm{c}}\widetilde{Q_{N_\mathrm{c}}}\|_{\mathfrak{S}_2(L^2_\#)} \\ &= \|(1-\Delta)^{1/2}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)}\|\widetilde{Q_{N_\mathrm{c}}}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)},\end{aligned}$$

since γ_{0,N_c} is a finite-rank orthogonal projector. Moreover, as the orbitals $(\phi_{i,N_\mathrm{c}})_{i=1,\cdots,N}$ are bounded in $H^1_\#(\Omega)$ uniformly in N_c, $\|(1-\Delta)^{1/2}\gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)}$ is also bounded uniformly in N_c. On top of that, noting that $\|(1-\Delta)^{-1/2}\| \leq 1$, we have

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

Therefore, we can use the estimate of Lemma 4.7 to conclude. □

From Lemma 4.4, and using the estimates of Lemmas 4.5, 4.7 and 4.8, we easily get estimate (37).

4.2.3 Proof of estimate (38)

If the perturbed density matrix were satisfying $\widetilde{\gamma_{N_\mathrm{c}}} \in \mathcal{K}$, we could deduce from (16) that the error $\widetilde{\mathcal{E}_{0,N_\mathrm{c}}} - \mathcal{E}_0$ would be non-negative and converge to zero as $\|(1-\Delta)^{1/2}(\widetilde{\gamma_{N_\mathrm{c}}}-\gamma_0)\|^2_{\mathfrak{S}_2(L^2_\#)}$ when N_c goes to infinity, yielding an improvement factor for the energy of order N_c^{-4}. However, as pointed out in Remark 3.1, $\widetilde{\gamma_{N_\mathrm{c}}}$ does not belong to $\mathcal{K}$ in general. We are going to show that the improvement factor for the energy is in fact of order N_c^{-2}.

We have $\widetilde{\mathcal{E}_{0,N_\mathrm{c}}}-\mathcal{E}_0 = \mathrm{Tr}\,(\gamma_{0,N_\mathrm{c}}\mathcal{H}\,\widetilde{\gamma_{N_\mathrm{c}}})-\mathrm{Tr}\,(\gamma_0\mathcal{H}\,\gamma_0)$. As $\mathrm{Tr}\,\left((\gamma_0)^2\right) = N$ and $\mathrm{Tr}\,(\gamma_{0,N_\mathrm{c}}\widetilde{\gamma_{N_\mathrm{c}}}) = \mathrm{Tr}\,(\gamma_{0,N_\mathrm{c}}^2) = N$, the energy difference can be written as follows

$$\begin{aligned}\widetilde{\mathcal{E}_{0,N_\mathrm{c}}} - \mathcal{E}_0 &= \mathrm{Tr}\,(\gamma_{0,N_\mathrm{c}}(\mathcal{H}-\epsilon_F)\widetilde{\gamma_{N_\mathrm{c}}}) - \mathrm{Tr}\,(\gamma_0(\mathcal{H}-\epsilon_F)\gamma_0) \\ &= \mathrm{Tr}\,((\gamma_{0,N_\mathrm{c}}-\gamma_0)(\mathcal{H}-\epsilon_F)(\widetilde{\gamma_{N_\mathrm{c}}}-\gamma_0)) + \mathrm{Tr}\,(\gamma_0(\mathcal{H}-\epsilon_F)(\widetilde{\gamma_{N_\mathrm{c}}}+\gamma_{0,N_\mathrm{c}}-2\gamma_0)) \\ &= \mathrm{Tr}\,((\gamma_{0,N_\mathrm{c}}-\gamma_0)(\mathcal{H}-\epsilon_F)(\widetilde{\gamma_{N_\mathrm{c}}}-\gamma_0)) + 2\mathrm{Tr}\,(\gamma_0(\mathcal{H}-\epsilon_F)(\gamma_{0,N_\mathrm{c}}-\gamma_0)) + \mathrm{Tr}\left(\gamma_0(\mathcal{H}-\epsilon_F)\gamma^{(1)}_{N_\mathrm{c}}\right).\end{aligned} \tag{54}$$

We now estimate each of these three terms. First, noting that $(\mathcal{H}-\epsilon_F) = -\gamma_0|\mathcal{H}-\epsilon_F| + (1-\gamma_0)|\mathcal{H}-\epsilon_F|$, using the triangle and the Cauchy–Schwarz inequalities, the fact that $|\mathcal{H}-\epsilon_F|, \gamma_0$ and $(1-\gamma_0)$ commute, and (13), we get

$$\begin{aligned}|\mathrm{Tr}\,((\gamma_{0,N_\mathrm{c}}-\gamma_0)(\mathcal{H}-\epsilon_F)(\widetilde{\gamma_{N_\mathrm{c}}}-\gamma_0))\,| &= \Big|\mathrm{Tr}\,((\gamma_{0,N_\mathrm{c}}-\gamma_0)(1-\gamma_0)|\mathcal{H}-\epsilon_F|(\widetilde{\gamma_{N_\mathrm{c}}}-\gamma_0)) \\ &\qquad - \mathrm{Tr}\,((\gamma_{0,N_\mathrm{c}}-\gamma_0)\gamma_0|\mathcal{H}-\epsilon_F|(\widetilde{\gamma_{N_\mathrm{c}}}-\gamma_0))\Big| \\ &\leq \Big(\||\mathcal{H}-\epsilon_F|^{1/2}(1-\gamma_0)(\gamma_{0,N_\mathrm{c}}-\gamma_0)\|_{\mathfrak{S}_2(L^2_\#)} \\ &\qquad + \||\mathcal{H}-\epsilon_F|^{1/2}\gamma_0(\gamma_{0,N_\mathrm{c}}-\gamma_0)\|_{\mathfrak{S}_2(L^2_\#)}\Big)\||\mathcal{H}-\epsilon_F|^{1/2}(\widetilde{\gamma_{N_\mathrm{c}}}-\gamma_0)\|_{\mathfrak{S}_2(L^2_\#)} \\ &\leq 2\||\mathcal{H}-\epsilon_F|^{1/2}(\gamma_{0,N_\mathrm{c}}-\gamma_0)\|_{\mathfrak{S}_2(L^2_\#)}\||\mathcal{H}-\epsilon_F|^{1/2}(\widetilde{\gamma_{N_\mathrm{c}}}-\gamma_0)\|_{\mathfrak{S}_2(L^2_\#)}.\end{aligned}$$

From (14) and (37), there exist $C \in \mathbb{R}_+$ and $N_\mathrm{c}^0 \in \mathbb{N}$, such that for all $N_\mathrm{c} \geq N_\mathrm{c}^0$,

$$|\mathrm{Tr}\,((\gamma_{0,N_\mathrm{c}}-\gamma_0)(\mathcal{H}-\epsilon_F)(\widetilde{\gamma_{N_\mathrm{c}}}-\gamma_0))\,| \leq CN_\mathrm{c}^{-2}\|(1-\Delta)^{1/2}(\gamma_{0,N_\mathrm{c}}-\gamma_0)\|^2_{\mathfrak{S}_2(L^2_\#)}. \tag{55}$$

Second, noting that for all $i = 1, \dots, N$, $\langle \phi_i^0 | \gamma_0 - \gamma_{0,N_\mathrm{c}} | \phi_i^0 \rangle \geq 0$, we get

$$\begin{aligned}
|\mathrm{Tr}\,(\gamma_0(\mathcal{H} - \epsilon_F)(\gamma_{0,N_\mathrm{c}} - \gamma_0)| &= \left| \sum_{i=1}^N (\lambda_i^0 - \epsilon_F) \langle \phi_i^0 | \gamma_{0,N_\mathrm{c}} - \gamma_0 | \phi_i^0 \rangle \right| \\
&\leq \max_{i=1,\dots,N} |\lambda_i^0 - \epsilon_F| \, \sum_{i=1}^N \langle \phi_i^0 | \gamma_0 - \gamma_{0,N_\mathrm{c}} | \phi_i^0 \rangle \\
&= \max_{i=1,\dots,N} |\lambda_i^0 - \epsilon_F| \, \left(N - \mathrm{Tr}\,(\gamma_{0,N_\mathrm{c}} \gamma_0) \right) \\
&= \frac{1}{2} \max_{i=1,\dots,N} |\lambda_i^0 - \epsilon_F| \, \|\gamma_0 - \gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)}^2.
\end{aligned}$$

From (25), there exist $C \in \mathbb{R}_+$ and $N_\mathrm{c}^0 \in \mathbb{N}$, such that for all $N_\mathrm{c} \geq N_\mathrm{c}^0$,

$$|\mathrm{Tr}\,((\gamma_{0,N_\mathrm{c}} - \gamma_0)(\mathcal{H} - \epsilon_F)(\widetilde{\gamma_{N_\mathrm{c}}} - \gamma_0))\,| \leq C N_\mathrm{c}^{-2} \|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_\mathrm{c}})\|_{\mathfrak{S}_2(L^2_\#)}^2. \tag{56}$$

Third, noting that for $i, j = 1, \dots, N$, $\langle \phi_{j,N_\mathrm{c}}^{(1)} | \phi_{i,N_\mathrm{c}}^0 \rangle = 0, \|\phi_i^0\|_{L^2_\#} = 1, \|\phi_{j,N_\mathrm{c}}\|_{L^2_\#} = 1$, and using (35) and the Cauchy–Schwarz inequality, we get

$$\begin{aligned}
\left| \mathrm{Tr}\,\left(\gamma_0 (\mathcal{H} - \epsilon_F) \gamma_{N_\mathrm{c}}^{(1)} \right) \right| &= \left| \sum_{i=1}^N (\lambda_i^0 - \epsilon_F) \langle \phi_i^0 | \gamma_{N_\mathrm{c}}^{(1)} | \phi_i^0 \rangle \right| \\
&= 2 \left| \sum_{i=1}^N \sum_{j=1}^N (\lambda_i^0 - \epsilon_F) \langle \phi_i^0 | \phi_{j,N_\mathrm{c}} \rangle \langle \phi_{j,N_\mathrm{c}}^{(1)} | \phi_i^0 \rangle \right| \\
&\leq 2 \max_{i=1,\dots,N} |\lambda_i^0 - \epsilon_F| \sum_{i=1}^N \sum_{j=1}^N \left| \langle \phi_i^0 | \phi_{j,N_\mathrm{c}} \rangle \langle \phi_{j,N_\mathrm{c}}^{(1)} | \phi_i^0 - \phi_{i,N_\mathrm{c}}^0 \rangle \right| \\
&\leq 2 \max_{i=1,\dots,N} |\lambda_i^0 - \epsilon_F| \sum_{i=1}^N \|\phi_i^0 - \phi_{i,N_\mathrm{c}}^0\|_{L^2_\#} \sum_{j=1}^N \|\phi_{j,N_\mathrm{c}}^{(1)}\|_{L^2_\#} \\
&\leq 2N \max_{i=1,\dots,N} |\lambda_i^0 - \epsilon_F| \|\Phi^0 - \Phi_{N_\mathrm{c}}^0\|_{L^2_\#} \left(\sum_{j=1}^N \|\phi_{j,N_\mathrm{c}}^{(1)}\|_{L^2_\#}^2 \right)^{1/2}.
\end{aligned}$$

Let us now estimate $\sum_{j=1}^N \|\phi_{j,N_\mathrm{c}}^{(1)}\|_{L^2_\#}^2$. Using (32)–(33) and noting that $\Pi_{N_\mathrm{c}}^\perp$ and $(1-\Delta)^{-1/2}$ commute, we get

$$\begin{aligned}
\sum_{i=1}^N \|\phi_{i,N_\mathrm{c}}^{(1)}\|_{L^2_\#}^2 &= \sum_{i=1}^N \|(-\frac{1}{2}\Delta - \lambda_{i,N_\mathrm{c}})^{-1} \Pi_{N_\mathrm{c}}^\perp \mathcal{V}_{N_\mathrm{c}}^\perp \phi_{i,N_\mathrm{c}}\|_{L^2_\#}^2 \\
&\leq \max_{i=1,\cdots,N} \|(-\frac{1}{2}\Delta - \lambda_{i,N_\mathrm{c}})^{-1}(1-\Delta)^{1/2} \Pi_{N_\mathrm{c}}^\perp\|^2 \sum_{i=1}^N \|(1-\Delta)^{-1/2} \mathcal{V}_{N_\mathrm{c}}^\perp \phi_{i,N_\mathrm{c}}\|_{L^2_\#}^2 \\
&\leq (1+N_\mathrm{c}^2)^{-1} \max_{i=1,\cdots,N} \|(-\frac{1}{2}\Delta - \lambda_{i,N_\mathrm{c}})^{-1}(1-\Delta)\|^2 \sum_{i=1}^N \|(1-\Delta)^{-1/2} \mathcal{V}_{N_\mathrm{c}}^\perp \phi_{i,N_\mathrm{c}}\|_{L^2_\#}^2 \\
&= (1+N_\mathrm{c}^2)^{-1} \max_{i=1,\cdots,N} \|(-\frac{1}{2}\Delta - \lambda_{i,N_\mathrm{c}})^{-1}(1-\Delta)\|^2 \|(1-\Delta)^{-1/2} \mathcal{V}_{N_\mathrm{c}}^\perp \gamma_{0,N_\mathrm{c}}\|_{\mathfrak{S}_2(L^2_\#)}^2.
\end{aligned}$$

Therefore, since $\max_{i=1,\cdots,N} \|(-\frac{1}{2}\Delta - \lambda_{i,N_\mathrm{c}})^{-1}(1-\Delta)\|^2$ is bounded uniformly in N_c, we deduce from (47)

that there exist $C \in \mathbb{R}_+$ and $N_{\rm c}^0 \in \mathbb{N}$, such that for all $N_{\rm c} \geq N_{\rm c}^0$,

$$\sum_{i=1}^{N} \|\phi_{i,N_{\rm c}}^{(1)}\|_{L^2_\#}^2 \quad \leq \quad C N_{\rm c}^{-2} \|(1-\Delta)^{1/2}(\gamma_{0,N_{\rm c}} - \gamma_0)\|_{\mathfrak{S}_2(L^2_\#)}^2. \tag{57}$$

From (22), (25) and (57), we obtain that there exist $C \in \mathbb{R}_+$ and $N_{\rm c}^0 \in \mathbb{N}$, such that for all $N_{\rm c} \geq N_{\rm c}^0$,

$$\left|\mathrm{Tr}\left(\gamma_0(\mathcal{H} - \epsilon_F)\gamma_{N_{\rm c}}^{(1)}\right)\right| \leq C N_{\rm c}^{-2} \|(1-\Delta)^{1/2}(\gamma_{0,N_{\rm c}} - \gamma_0)\|_{\mathfrak{S}_2(L^2_\#)}^2. \tag{58}$$

Putting together (54), (55), (56), and (58), we obtain estimate (38).

5 Numerical results

We present in this section some results to illustrate the statements of Theorem 4.1 for several eigenvalue clusters and potentials with different regularities. First note that the quantities $\phi_{j,N_{\rm c}}^{(1)}$ defined in (32) are easily computable. Indeed, the operator $\left(-\frac{1}{2}\Delta - \lambda_{j,N_{\rm c}}\right)$ is diagonal in plane-wave bases, hence very easy to invert. Moreover, only two FFT's are needed to compute the residual or $\mathcal{V}_{N_{\rm c}}^{\perp}\phi_{j,N_{\rm c}}$ on a larger grid, via a product in the physical space. However we focus here mainly on the convergence rate improvement, and not in the low computational cost of the method, which has been demonstrated for the nonlinear problem of Kohn–Sham equations in [7].

In all what follows, we consider a domain $\Omega = (0,10)^3$ in atomic units (a.u.). The computed solutions are compared to a reference solution, computed in a very large basis with a kinetic energy cutoff $E_{\rm ref} = 800$ a.u., which corresponds to a discretization parameter $N_{\rm ref} \simeq 58.5$, and 382,323 Fourier basis functions. In each case, we denote the reference energy by $\mathcal{E}_0$ and the reference density matrix by γ_0.

The coarse solutions are computed in a basis with cutoff E_c and corresponding N_c, and have energy $\mathcal{E}_{0,N_{\rm c}}$ and density matrix $\gamma_{0,N_{\rm c}}$. In order to avoid errors coming from the size of the finite basis used for the computation of the corrections, we compute the post-processed solutions in the same basis as the reference solution, *i.e.* in a basis with energy cutoff $E_{\rm ref}$. Note that the components of the orbitals in the coarse basis are not modified by the post-processing. One only needs to compute the coefficients corresponding to Fourier modes with kinetic energy larger than E_c.

The implementation is based on KSSOLV [20], a Matlab library for solving Kohn–Sham equations, which we use here to solve the linear eigenvalue problem (29).

The tested potentials denoted by $\mathcal{V}_s$ are defined by their Fourier coefficients as

$$\widehat{\mathcal{V}}_{s,\mathbf{0}} = 0, \quad \text{and} \quad \forall \mathbf{k} \in \mathcal{R}^* \backslash \{\mathbf{0}\}, \quad \widehat{\mathcal{V}}_{s,\mathbf{k}} = -\frac{c_s}{|\mathbf{k}|^{2s}},$$

where $s \in [1,2]$ is a regularity parameter, and c_s is a multiplicative constant. For $s > 3/2$, the potential $\mathcal{V}_s$ is smooth enough to verify Assumption 2.2, hence we expect to observe the improvement in the convergence rate given in Theorem 4.1. For $s \leq 3/2$, the potential does not verify Assumption 2.2, but we can nevertheless compute post-processed eigenfunctions and eigenvalues. As we will see, it actually still yields an improvement on the energy and the density matrix. Note that for $s = 1$, this potential has the same regularity as the Coulomb potential.

The lowest eigenvalue of the Hamiltonian $-\frac{1}{2}\Delta + \mathcal{V}$ is simple. For all the tested potentials, there are gaps between the $5^{\rm th}$ and $6^{\rm th}$ eigenvalues and the $10^{\rm th}$ and $11^{\rm th}$ eigenvalues. Therefore, in the following, we present the results of the post-processing method for clusters including one, five, and ten eigenvalues. This guarantees that the gap Assumption 2.1 is satisfied. Let us remind that we consider the cluster of lowest eigenvalues of the Hamiltonian.

In Subsection 5.1, we show how the post-processing procedure decreases both the energy error $\mathcal{E}_{0,N_{\rm c}} - \mathcal{E}_0$ and the energy norm $\|(1-\Delta)^{1/2}(\gamma_{0,N_{\rm c}} - \gamma_0)\|_{\mathfrak{S}_2(L^2_\#)}$ of the density matrix error in the case of a potential with regularity parameter $s = 2$ and a cluster of five eigenvalues. In Subsection 5.2, we study the convergence

rate improvement of both the energy and the density matrix for different clusters of eigenvalues, still in the case of a potential with regularity parameter $s = 2$. Finally, we study in Subsection 5.3 the influence of the potential regularity s on the convergence rate improvement for the energy and the density matrix, in the case of the cluster composed of the lowest five eigenvalues.

5.1 Convergence of the density matrix and the energy

We consider the potential $\mathcal{V}_2$ with Fourier coefficients

$$\widehat{\mathcal{V}}_{2,\mathbf{0}} = 0, \quad \text{and} \quad \forall \mathbf{k} \in \mathcal{R}^* \backslash \{\mathbf{0}\}, \quad \widehat{\mathcal{V}}_{2,\mathbf{k}} = -\frac{0.01}{|\mathbf{k}|^4}.$$

For all energy cutoffs E_c between 10 and 200 a.u. by steps of 10, we compute the lowest five eigenvalues and eigenvectors of the discrete Hamiltonian. We build the discrete density matrix γ_{0,N_c} as in (19), and compute the discrete energy $\mathcal{E}_{0,N_c}$ (20). Then, we apply the post-processing as described in Section 3 and we compute the post-processed density matrix $\widetilde{\gamma_{N_c}}$ as well as the perturbed energy $\widetilde{\mathcal{E}_{0,N_c}}$. For the potential, we choose such a small multiplicative constant ($c_s = 0.01$) to better observe the asymptotic regime numerically within the range of tested cutoffs E_c.

As we can see on the top part of Figure 1, the energy error between the post-processed energy and the reference energy $\widetilde{\mathcal{E}_{0,N_c}} - \mathcal{E}_0$ is 5 to 50 times smaller than the energy error between the coarse energy and the reference energy $\mathcal{E}_{0,N_c} - \mathcal{E}_0$. More precisely, the energy error is reduced by a factor of about 5 for small values of N_c and up to 50 for large values of N_c.

We observe a similar behavior for the density matrix error on the bottom part of Figure 1. Indeed, the Hilbert–Schmidt norm of the difference between the reference and the coarse density matrices $\|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_c})\|_{\mathfrak{S}_2(L^2_\#)}$ is 5 to 50 times larger than the error between the reference and the post-processed density matrix $\|(1-\Delta)^{1/2}(\gamma_0 - \widetilde{\gamma_{N_c}})\|_{\mathfrak{S}_2(L^2_\#)}$.

5.2 Comparison between different eigenvalue clusters

We now consider three different eigenvalue clusters composed of one, five and ten eigenvalues, still with the same potential $\mathcal{V}_2$. The three corresponding gaps are respectively equal to $8.84 \cdot 10^{-1}$, $1.80 \cdot 10^{-1}$ and $3.42 \cdot 10^{-1}$.

For these three clusters, we compute a reference solution, and then we compute discrete solutions within cutoffs E_c varying between 10 and 200 a.u. On the top of Figure 2, we plot the ratio between the energy error with post-processing and without post-processing $\frac{\widetilde{\mathcal{E}_{0,N_c}} - \mathcal{E}_0}{\mathcal{E}_{0,N_c} - \mathcal{E}_0}$ for the three different cases. According to Theorem 4.1, this ratio should at least decrease as N_c^{-2} in the asymptotic regime of large N_c's. This is approximately satisfied with the decay $N_c^{-1.74}$ observed in the numerical simulations. The difference between the expected and observed rate might come from a pre-asymptotic effect. Likewise, the ratio $\frac{\|(1-\Delta)^{1/2}(\gamma_0 - \widetilde{\gamma_{N_c}})\|_{\mathfrak{S}_2(L^2_\#)}}{\|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_c})\|_{\mathfrak{S}_2(L^2_\#)}}$ of the energy norms of the density matrix errors behaves in the asymptotic regime like $N_c^{-1.8}$ for one and five eigenvalues, and N_c^{-2} for the cluster with ten eigenvalues, as shown on the bottom of Figure 2.

5.3 Comparison of different regularities

Lastly, we perform the post-processing in the case of the cluster containing the lowest five eigenvalues with four potentials having different regularity coefficients. More precisely, we consider the potentials $\mathcal{V}_1, \mathcal{V}_{1.25}, \mathcal{V}_{1.5}$, and $\mathcal{V}_2$, again taking $c_s = 0.01$. In theory, the potential $\mathcal{V}_2$ satisfies Assumption 2.2, the potential $\mathcal{V}_{1.5}$ is just at the limit, and $\mathcal{V}_{1.25}, \mathcal{V}_1$ do not satisfy the assumption. It is nevertheless possible to compute the post-processed energy and density matrix for each of these potentials. Numerically, we observe an improved convergence rate both for the energy error ratio (on the top of Figure 3) and for the density

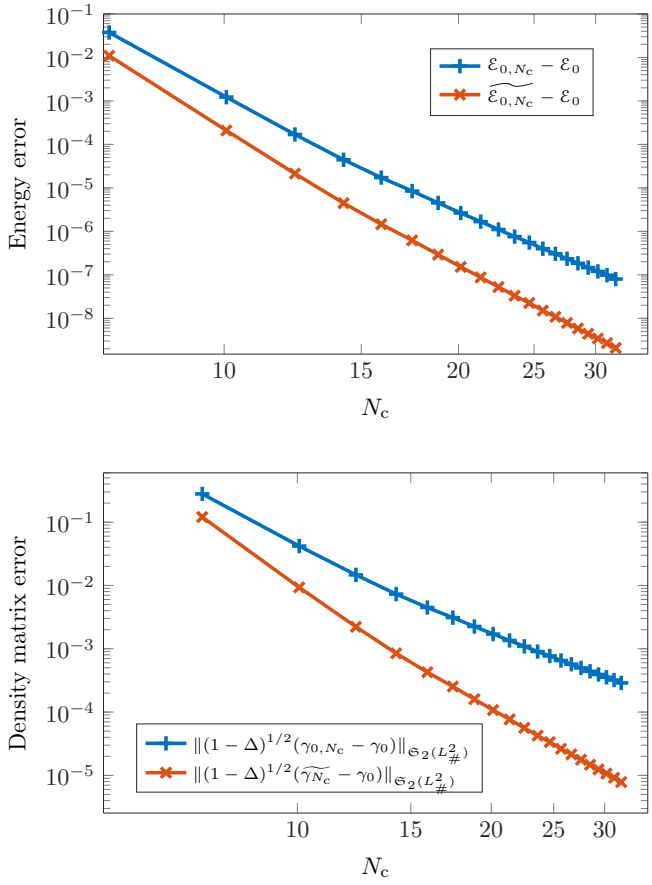

Figure 1: Top: plot of the energy errors $\widetilde{\mathcal{E}_{0,N_c}} - \mathcal{E}_0$ and $\mathcal{E}_{0,N_c} - \mathcal{E}_0$ for energy cutoffs E_c between 10 and 200 a.u. Bottom: plot of the density matrix errors $\|(1-\Delta)^{1/2}(\gamma_0 - \widetilde{\gamma_{N_c}})\|_{\mathfrak{S}_2(L^2_\#)}$ and $\|(1-\Delta)^{1/2}(\gamma_0 - \gamma_{0,N_c})\|_{\mathfrak{S}_2(L^2_\#)}$ for energy cutoffs between 10 and 200 a.u. This corresponds to values of N_c between 7 and 31.

matrix error ratio (on the bottom of Figure 3) for all potentials, which slowly and monotonically decreases as the regularity of the potential decreases. Hence, this post-processing method seems to yield an improvement also in a larger setting than what is covered by the proof of this article. Also, note that for the potentials $\mathcal{V}_{1.25}$ and $\mathcal{V}_1$ with low regularity, the tested E_c are far from convergence, so the asymptotic regime might not have been reached.

Appendix: proof of Lemma 2.3

We start by proving (22). Denoting by M_{N_c} the $N \times N$ overlap matrix with entries $(M_{N_c})_{i,j} = \langle \phi^0_{i,N_c} | \phi^0_j \rangle$, we have

$$\|\gamma_{0,N_c} - \gamma_0\|^2_{\mathfrak{S}_2(L^2_\#)} = 2\left(N - \sum_{i,j=1}^{N} |\langle \phi^0_{j,N_c} | \phi^0_i \rangle|^2\right) = 2\left(N - \mathrm{Tr}\left(M_{N_c} M_{N_c}^T\right)\right)$$

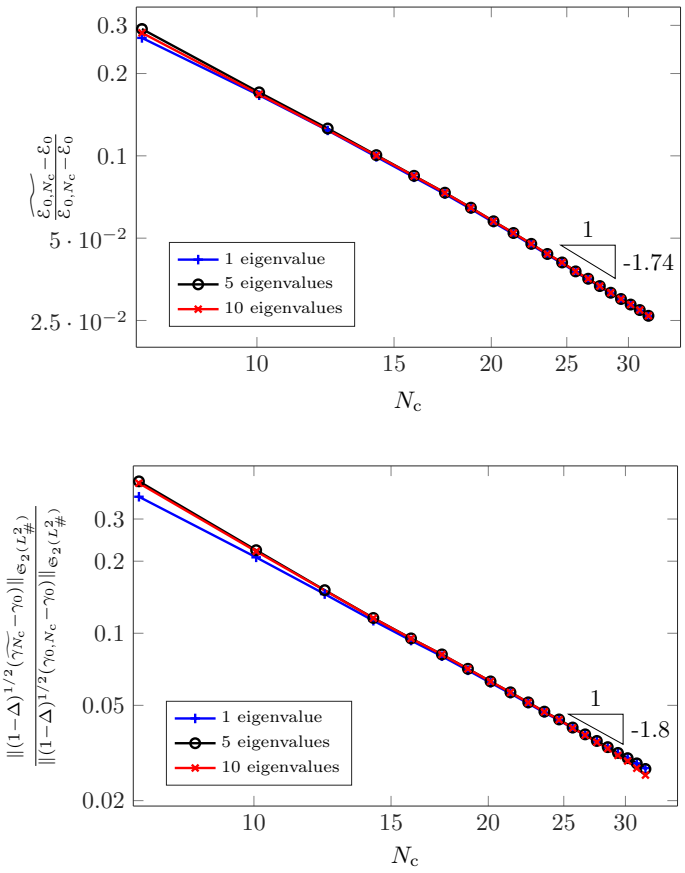

Figure 2: Plots of the energy error ratio (top) and the density matrix error ratio (bottom) for three different clusters of eigenvalues (1, 5 and 10 eigenvalues) with a potential with regularity coefficient $s = 2$. The convergence rates are computed with the largest 10 values of N_c respectively yielding to $N_c^{-1.73}$, $N_c^{-1.74}$, $N_c^{-1.74}$ (top), and $N_c^{-1.79}$, $N_c^{-1.84}$, $N_c^{-2.01}$ (bottom).

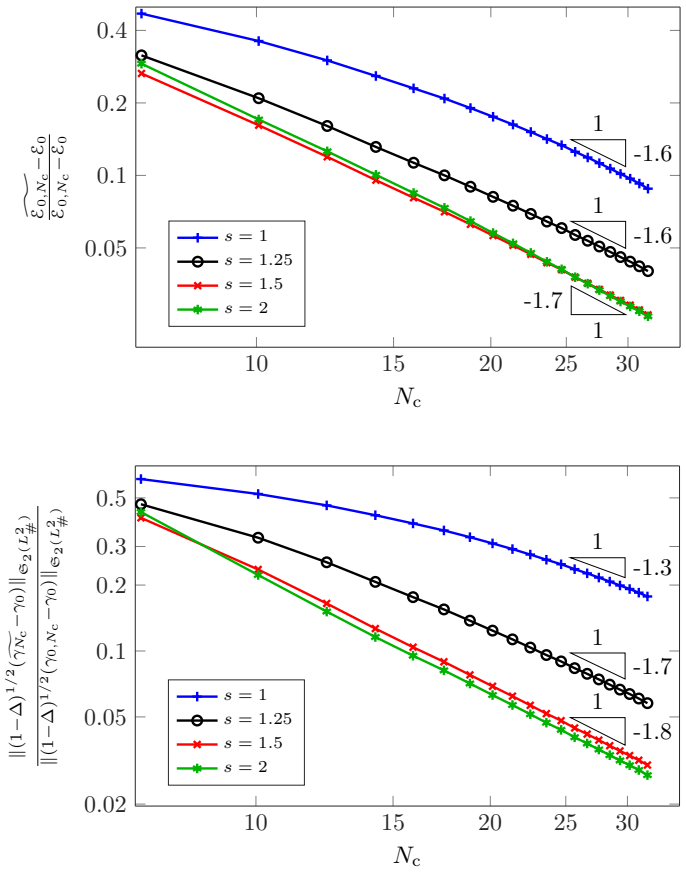

Figure 3: Plots of the energy error improvement (top) and the density matrix error improvement (bottom) for four different regularities for the potential: $s = 1, 1.25, 1.5, 2$, cluster of five eigenvalues. The convergence rates are computed with the largest 10 values of N_c respectively yielding to $N_c^{-1.59}$, $N_c^{-1.58}$, $N_c^{-1.67}$, $N_c^{-1.74}$ (top), and $N_c^{-1.29}$, $N_c^{-1.69}$, $N_c^{-1.81}$, $N_c^{-1.84}$ (bottom).

and

$$\|\Phi^0_{N_c} - \Phi^0\|^2_{L^2_\#} = 2\left(N - \sum_{i=1}^{N} \langle \phi^0_{i,N_c} | \phi^0_i \rangle\right) = 2\left(N - \mathrm{Tr}\left(M_{N_c}\right)\right).$$

We therefore have to show that

$$2N - 2\mathrm{Tr}\left(M_{N_c}\right) \leq 2N - 2\mathrm{Tr}\left(M_{N_c} M^T_{N_c}\right) \leq 2(2N - 2\mathrm{Tr}\left(M_{N_c}\right)).$$

Since $\Phi^0_{N_c}$ belongs to $\mathcal{M}^{\Phi^0}$, we have from [3, Lemma 4.3]

$$M_{N_c} = M^T_{N_c} = (M_{N_c} M^T_{N_c})^{1/2} \quad \text{and} \quad 0 \leq M_{N_c} \leq 1.$$

Hence,

$$\mathrm{Tr}\left(M_{N_c} M^T_{N_c}\right) = \mathrm{Tr}\left(M^2_{N_c}\right) \leq \mathrm{Tr}\left(M_{N_c}\right),$$

from which we deduce the left inequality in (22). The right inequality in (22) holds since

$$2N - 4\mathrm{Tr}\left(M_{N_c}\right) + 2\mathrm{Tr}\left(M_{N_c} M^T_{N_c}\right) = 2N - 4\mathrm{Tr}\left(M_{N_c}\right) + 2\mathrm{Tr}\left(M^2_{N_c}\right) = 2\mathrm{Tr}\left(\left(M_{N_c} - I_N\right)^2\right) \geq 0.$$

Let us now show (23). From [3, Theorem 4.2], there exist two constants $0 < c \leq C < \infty$ and $N^0_c \in \mathbb{N}$ such that for all $N_c \geq N^0_c$,

$$c\|(1-\Delta)^{1/2}(\Phi^0_{N_c} - \Phi^0)\|^2_{L^2_\#} \leq \mathcal{E}_{0,N_c} - \mathcal{E}_0 \leq C\|(1-\Delta)^{1/2}(\Phi^0_{N_c} - \Phi^0)\|^2_{L^2_\#}.$$

Hence, using (16) finishes the proof.

Acknowledgements

This work was supported by the ANR project MANIF Mathematical and numerical issues in first-principle molecular simulation. Part of this work has been supported from French state funds managed by the CalSimLab LABEX and the ANR within the Investissements d'Avenir program (reference ANR-11-LABX-0037-01). MV has also received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 647134 GATIPOR). BS acknowledges the funding from the German Academic Exchange Service (DAAD) from funds of the Bundesministeriums fr Bildung und Forschung (BMBF) for the project Aa-Par-T (Project-ID 57317909). YM and EC acknowledge the funding from the PICS- CNRS and the PHC PROCOPE 2017 (Project No. 37855ZK).

References

[1] M. Born and R. Oppenheimer, *Zur Quantentheorie der Molekeln*, Annalen der Physik, 389 (1927), pp. 457–484.

[2] E. Cancès, R. Chakir, L. He, and Y. Maday, *Two-grid methods for a class of nonlinear elliptic eigenvalue problems*, IMA J. Numer. Anal., 38 (2018), pp. 605–645.

[3] E. Cancès, R. Chakir, and Y. Maday, *Numerical analysis of the planewave discretization of some orbital-free and Kohn-Sham models*, ESAIM: Mathematical Modelling and Numerical Analysis, 46 (2012), pp. 341–388.

[4] E. Cancès, M. Defranceschi, W. Kutzelnigg, C. Le Bris, and Y. Maday, *Computational quantum chemistry: A primer*, in Handbook of Numerical Analysis, vol. 10, 2003, pp. 3–270.

[5] E. Cancès, A. Deleurence, and M. Lewin, *A new approach to the modeling of local defects in crystals: The reduced Hartree–Fock case*, Communications in Mathematical Physics, 281 (2008), pp. 129–177.

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

[7] ——, *A perturbation-method-based post-processing for the planewave discretization of Kohn–Sham models*, Journal of Computational Physics, 307 (2016), pp. 446–459.

[8] F. Chatelin, *Spectral Approximation of Linear Operators*, Computer science and applied mathematics, Academic Press, 1983.

[9] R. M. Dreizler and E. K. U. Gross, *Density Functional Theory*, Springer Berlin Heidelberg, Berlin, Heidelberg, 1990.

[10] G. Dusson, *Post-processing of the planewave approximation of Schrödinger equations. Part II: Kohn-Sham models, (in preparation)*. 2018.

[11] T. Helgaker, P. Jørgensen, and J. Olsen, *Molecular Electronic Structure Theory*, John Wiley & Sons, LTD, Chichester, 2000.

[12] M. Hoffmann-Ostenhof and T. Hoffmann-Ostenhof, *"Schrödinger inequalities" and asymptotic behavior of the electron density of atoms and molecules*, Phys. Rev. A, 16 (1977), pp. 1782–1785.

[13] T. Kato, *Perturbation Theory for Linear Operators*, Classics in Mathematics, Springer Berlin Heidelberg, Berlin, Heidelberg, 1976.

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

[15] W. A. Lester, *Recent Advances in Quantum Monte Carlo Methods*, vol. 2 of Recent Advances in Computational Chemistry, World Scientific, 1997.

[16] W. A. Lester, S. M. Rothstein, and S. Tanaka, *Recent Advances in Quantum Monte Carlo Methods - Part II*, vol. 2 of Recent Advances in Computational Chemistry, World Scientific, 2002.

[17] Y. Maday and G. Turinici, *Error bars and quadratically convergent methods for the numerical simulation of the Hartree-Fock equations*, Numerische Mathematik, 94 (2003), pp. 739–770.

[18] M. Reed and B. Simon, *Methods of Modern Mathematical Physics, Vol. 1: Functional Analysis*, Academic Press, 1981.

[19] J. Xu and A. Zhou, *A two-grid discretization scheme for eigenvalue problems*, Mathematics of Computation, 70 (1999), pp. 17–26.

[20] C. Yang, J. C. Meza, B. Lee, and L.-W. Wang, *KSSOLV–a MATLAB toolbox for solving the Kohn-Sham equations*, ACM Transactions on Mathematical Software, 36 (2009), pp. 1–35.