A variational framework for spectral approximations of Kohn-Sham Density Functional Theory

X. Wang[1], T. Blesgen[2], K. Bhattacharya[1], and M. Ortiz[1]

[1]Division of Engineering and Applied Sciences, California Institute of Technology, USA
[2]Department of Mathematics, University of Applied Sciences, Bingen, Germany

Abstract

Kohn-Sham density functional theory (K-S DFT) is widely used to study the electronic structure of materials. The central difficulty in K-S DFT involves the solution of a non-linear eigenvalue problem. This non-linear problem is solved numerically by the self-consistent field method, a fixed point iteration approach, which yields linear eigenvalue problems. Typical solution of the linear eigenvalue problem is to diagonalize the matrix of the differential operator, the Hamiltonian of the system. There have been approximate solutions to K-S DFT that exploit spectral theory of self-adjoint operators, known as the density matrix expansion methods. These methods can avoid diagonalization of the Hamiltonian matrix. They are increasingly used to study the linearized problem because of their computational efficiency. Although these approximations have been verified numerically, the relationship between these approximations of the linearized problem and the original non-linear problem remain incompletely understood. Further, these methods assume smoothness that give rise to errors in conductors. In this paper, we reformulate K-S DFT as a nested variational problem that enables density matrix expansions. We introduce a new approximation, called the spectral binning discretization, which does not require smoothness. We show convergence with respect to both spectral binning discretization and with spatial discretization.

1 Introduction

The wave formulation of quantum mechanics proposed by Erwin Schrödinger in 1926 can be used in theory to quantitatively study the electronic structure of materials. However, it is limited to only a handful of electrons due to the high dimensionality of the resulting partial-differential equation. An approximate formulation, called the Hartree-Fock (H-F) method, was introduced in 1930 to reduce the dimensionality of the wave formulation. This reduction is achieved by variationally minimizing the energy over the set of Slater–determinant combinations of independent-electron orbitals, resulting in a non-linear eigenvalue problem in three dimensions [22]. The solution of the H-F equations is nevertheless cumbersome.

Density functional theory (DFT) developed by Kohn and Hohenberg in 1964 lay the foundation for the majority of approximate quantum mechanical methods used today. The Kohn-Hohenberg theorem provides a one-to-one correspondence between the ground-state electron density and the ground-state energy; thereby proving the existence of an ground-state energy functional that depends only the ground-state electron density. However, the exact form of the energy functional is unknown. Shortly after, Kohn and Sham introduced the Kohn-Sham density functional theory (K-S DFT), which provided an approximate energy functional by introducing explicit models for the kinetic energy and exchange-correlation functionals of the electron density. The resulting K-S problem is also a non-linear eigenvalue problem in three dimensions, but its non-linearity is less computational intensive than the H-F formulation. K-S DFT is widely used to study the electronic structure of materials ranging from molecules, macromolecules to crystalline solids [22].

The non-linear Kohn-Sham equations are solved by fixed point iterations or the self-consistent field method. In each step of the iteration, one calculates the sum of the N lowest eigenvalues of the linearized Hamiltonian and the resulting electron density, where N denotes the number of electrons in the system of interest. If this procedure is carried out using direct diagonalization, the computational effort scales to the third power ($O(N^3)$) with respect to the number of electrons N in the system. This scaling limits K-S DFT calculations routinely to systems with only a few hundred to thousand electrons. With pseudo-potential approximations, where the core electrons are lumped with the nuclei, K-S DFT calculations can be done for a few hundred atoms.

However, hundreds of atoms are not sufficient to study materials with defects or complex macromolecules. Defects often occur in real materials in parts per million concentrations. Therefore, a number of linear-scaling algorithms, where the computational cost scales linearly ($O(N)$), have been developed (see [19, 7] for a review, and [5, 20, 29, 30, 26, 32] for specific implementations). The key idea behind these methods is to introduce the *density matrix*,

$$\gamma = \sum_{1 \le i \le N} \psi_i \otimes \psi_i,$$

where ψ_i denotes the eigenvectors corresponding to the lowest eigenvalues of the linearized Hamiltonian H. It follows then from spectral theory (cf. for example [24]) that

$$\gamma = f(H), \tag{1}$$

where the occupancy function $f : \mathbb{R} \to \mathbb{R}$ is

$$f(\lambda) = \begin{cases} 1, & \text{if } \lambda \le \lambda_N, \\ 0, & \text{otherwise.} \end{cases}$$

It is common at this stage to regularize f by introducing a temperature σ and to replace it with the Fermi-Dirac distribution,

$$f^{\mathrm{FD}}(\lambda) = \frac{1}{1 + \exp \frac{\lambda - \lambda_N}{\sigma}}. \tag{2}$$

Note that the regularization can be made exact in insulators/semiconductors where there is a non-zero gap between λ_N and λ_{N+1}, but is only approximate in conductors.

The main idea behind the linear-scaling methods is to expand f^{FD} using polynomials, rational functions, etc. In some methods (e.g., [5]), a proper choice of spatial discretization leads to a fast decay of the off-diagonal elements of γ. Therefore, one truncates γ to obtain a banded matrix. The expansion can then be carried out at linear-computational cost. In the recently introduced linear-scaling spectral Gauss quadrature (LSSGQ) method, [27], one takes advantage of the sparsity of the Hamiltonian matrix as result of an appropriate basis set in the Lanczos iteration to obtain linear-scaling without truncating the density matrix.

All these linear-scaling methods, with or without truncation, have two significant shortcomings. Firstly, they approximate the density matrix of the linearized problem obtained from an iteration of the self-consistent scheme. There are results that establish the convergence in the linearized eigenvalue problem, [29]. However, to our knowledge, there is no rigorous study showing convergence of this approach to the original Kohn-Sham equations. Secondly, they involve the regularization of the occupancy function. These two shortcomings motivate the work presented in this paper.

The original Kohn-Sham equations may be written as a variational principle over the density matrices, trace-class operators. In fact, Anantharaman and Cancès, [2], have done so rigorously and proved existence of solutions to this variational problem even in an unbounded domain. However, the functional is not amenable to the application of simple spectral representation.

In this work, we reformulate the variational principle to enable simple spectral representation. The main idea is to use duality in the exchange-correlation functional, thereby converting the original formulation to a nested variational problem. The resulting functional is linear in the density matrix and thus amenable to simple spectral representation.

We then introduce a new class of operator approximations, spectral binning discretization, using simple functions on the spectrum that enables an accurate representation of the occupancy function without regularization. We show convergence with respect to combined spatial and spectral discretizations.

While spectral binning discretization provides an exact representation, a practical and efficient numerical implementation remains an open issue. As a first step, we study a linear one-dimensional model problem that has been used as a step towards Kohn-Sham equations, and show that spectral binning discretization is potentially very attractive.

This paper is organized as follows. Section 2 recalls the Kohn-Sham density functional theory and reformulates it as a nested variational problem. Section 3 collects the main theorems of existence and convergence. Section 4 presents the proof of the existence of minimizers. Section 5 describes spatial and spectral discretization. Section 6 presents the proof of convergence with combined spatial and spectral discretization. Section 7 is a numerical demonstration of spectral binning in a one-dimensional linear model-problem.

2 Kohn-Sham density functional theory

For simplicity, we restrict ourselves to closed-shell, spin-unpolarized systems. We also restrict ourselves to an open and bounded subset Ω of $\mathbb{R}^3$. This is an important restriction since the formulation in $\mathbb{R}^3$ introduces non-trivial difficulties. We also restrict ourselves to the local density approximation (LDA) for the exchange-correlation. Finally we make, as common in this subject, the Born-Oppenheimer hypothesis that the atomic nuclei are classical. So we hold the nuclei fixed in the rest of the section.

We start with the operator formulation due to Anantharaman and Cancès, [2]. The connection to the traditional orbital formulation is given in Appendix B.

2.1 Operator formulation

Let $\mathcal{V}=\mathcal{W}_0^{1,2}(\Omega)$, $\mathcal{H}=\mathcal{L}^2(\Omega)$ and $\mathfrak{S}_1$ be the vector space of self-adjoint, trace-class operators on $\mathcal{H}$,

$$\mathfrak{S}_1 = \{\gamma \in \mathcal{S}(\mathcal{H}) : \mathrm{Tr}(|\gamma|) < \infty\}, \tag{3}$$

where $|\gamma| \equiv \sqrt{\gamma\gamma^*}$. $\mathfrak{S}_1$ is a separable Banach space [4]. Within $\mathfrak{S}_1$, we can introduce the space

$$\mathcal{X} = \{\gamma \in \mathfrak{S}_1 : |\nabla|\gamma|\nabla| \in \mathfrak{S}_1\},$$

and the constrained set of admissible reduced one-particle density operators,

$$\mathcal{K}_N = \{\gamma \in \mathcal{X} \,:\, 0 \leq \gamma \leq 1, \mathrm{Tr}(\gamma) = N\}. \tag{4}$$

Remark 2.1 *As stated in [2], for every $\gamma \in \mathcal{K}_N$, we have the canonical representation in the continuous r basis,*

$$\gamma(\mathbf{r}, \mathbf{r}') = \sum_{i=1}^{\infty} 2\alpha_i \xi_i(\mathbf{r})\xi_i(\mathbf{r}'), \tag{5}$$

where $\xi_i \in \mathcal{V}$ for all $i \in \mathbb{N}$, the factor of 2 simply accounting for spin unpolarization, and

$$0 \le \alpha_i \le 1, \qquad \int_\Omega \xi_i(\mathbf{r})\xi_j(\mathbf{r})\,\mathrm{d}\mathbf{r} = \delta_{ij}, \qquad \sum_{i=1}^{\infty} 2\alpha_i = N.$$

We can define the electron density for every $\gamma \in \mathcal{K}_N$ as

$$\rho_\gamma(\mathbf{r}) = \gamma(\mathbf{r}, \mathbf{r}).$$

We consider a system of M atoms with nuclei located at $\{\mathbf{R}\} = \{\mathbf{R}_1, \dots, \mathbf{R}_M\} \subset \Omega$ and nuclear charges $Z_1, \dots, Z_M$. We now follow Anantharaman and Cancès, [2], and define the extended Kohn-Sham energy functional $E^{\mathrm{EKS}} : \mathcal{K}_N \to \mathbb{R}$ as

$$E^{\mathrm{EKS}}(\gamma) = T_{\mathrm{s}}(\gamma) + E_{\mathrm{H}}(\rho_\gamma) + E_{\mathrm{ext}}(\rho_\gamma) + E_{\mathrm{ZZ}} + E_{\mathrm{xc}}(\rho_\gamma), \tag{6}$$

where T_{S} is the kinetic energy of the non-interacting electrons,

$$T_{\mathrm{S}}(\gamma) = \mathrm{Tr}\left(-\frac{1}{2}\Delta\gamma\right),$$

E_{H} is the Hartree energy representing the classical electrostatic repulsion energy for a given electron density,

$$E_{\mathrm{H}}(\rho_\gamma) = \frac{1}{2}\int_\Omega\int_\Omega \frac{\rho_\gamma(\mathbf{r})\rho_\gamma(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}\,\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{r}', \tag{7}$$

E_{ext} is the interaction energy between the nuclear charges and the electrons,

$$E_{\mathrm{ext}}(\rho_\gamma) = \int_\Omega \rho_\gamma(\mathbf{r}) V_{\mathrm{ext}}(\mathbf{r}, \{\mathbf{R}\})\,\mathrm{d}\mathbf{r} = \int_\Omega \rho_\gamma(\mathbf{r}) \left(\sum_{1 \le I \le M} \frac{Z_I}{|\mathbf{R}_I - \mathbf{r}|}\right) \mathrm{d}\mathbf{r}, \tag{8}$$

E_{ZZ} is the classical electrostatic repulsion energy due to the nuclear charges,

$$E_{\mathrm{ZZ}} = \frac{1}{2} \sum_{1 \le I \le J \le M} \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|}, \tag{9}$$

and $E_{\mathrm{xc}}(\rho_\gamma)$ is the exchange-correlation energy that is split into two terms (cf. [23]),

$$E_{\mathrm{xc}}(\rho_\gamma) = E_{\mathrm{x}}(\rho_\gamma) + E_{\mathrm{c}}(\rho_\gamma) = \int_\Omega h(\rho_\gamma)\,\mathrm{d}\mathbf{r}, \tag{10}$$

with an exchange term,

$$E_{\mathrm{x}}(\rho_\gamma) = -\frac{3}{4}\left(\frac{6}{\pi}\right)^{1/3} \int_\Omega \rho_\gamma^{4/3}(\mathbf{r})\,\mathrm{d}\mathbf{r},$$

and a correlation term,

$$E_{\mathrm{c}}(\rho_\gamma) = \int_\Omega \epsilon_{\mathrm{c}}(\rho_\gamma(\mathbf{r}))\rho_\gamma(\mathbf{r})\,\mathrm{d}\mathbf{r},$$

where ϵ_{c} is taken from [23]. The connection of this formulation to the traditional formulation is in Appendix A.

The ground-state energy of the extended Kohn-Sham energy functional is

$$\varepsilon_{\mathrm{GS}}^{\mathrm{EKS}} = \inf_{\gamma \in \mathcal{K}_N} E^{\mathrm{EKS}}(\gamma).$$

The existence of minimizers of the extended Kohn-Sham energy functional has been shown in [2].

2.2 Reformulation

The above formulation of the extended K-S energy functional is not amenable to spectral discretization algorithms because of the non-linearity in the terms E_{H} and E_{xc}. To overcome this, we reformulate these terms as follows.

2.2.1 Electrostatics

We reformulate the electrostatic terms by writing them as the solution to a Helmholtz problem (cf., e.g., [3, 28]). We approximate the nuclear charges at a given atomic site $\mathbf{R}_i$ by a regularized and bounded nuclear charge distribution $-Z_i f_{\mathbf{R}_i}(\mathbf{r})$ with compact support on a small ball centered at $\mathbf{R}_i$ satisfying

$$\int_\Omega f_{\mathbf{R}_i}(\mathbf{r})\,\mathrm{d}\mathbf{r} = 1.$$

We can then rewrite the electrostatic terms as the variational problem

$$\begin{aligned} &E_{\mathrm{H}}(\rho_\gamma) + E_{\mathrm{ext}}(\rho_\gamma) + E_{\mathrm{ZZ}} \\ &\quad = \sup_{\phi \in \mathcal{V}} \left\{ -C_S \int_\Omega |\nabla \phi(\mathbf{r})|^2\,\mathrm{d}\mathbf{r} + \int_\Omega (b(\mathbf{r},\{\mathbf{R}\}) + \rho_\gamma(\mathbf{r}))\phi(\mathbf{r})\,\mathrm{d}\mathbf{r} \right\} + C_{\mathrm{self}}, \end{aligned}$$

where

$$b(\mathbf{r},\{\mathbf{R}\}) = \sum_{i=1}^{M} Z_i f_{\mathbf{R}_i}(\mathbf{r}),$$

$C_S > 0$ is a constant depending on the spatial dimension S (e.g. $C_S = \frac{1}{8\pi}$ for $S = 3$); C_{self} is an inessential constant that depends only on the regularization $f_{\mathbf{R}_i}$ and is independent of ρ_γ and $\{\mathbf{R}\}$.

To clarify the dependence of the electrostatic terms on γ, we introduce an unbounded local operator,

$$\Phi(\mathbf{r},\mathbf{r}') = \phi(\mathbf{r})\delta(\mathbf{r},\mathbf{r}'), \tag{11}$$

and use its coordinate representation so that

$$\mathrm{Tr}(\Phi\gamma) = \int_\Omega \phi(\mathbf{r})\rho_\gamma(\mathbf{r})\,\mathrm{d}\mathbf{r}.$$

The Coulomb energy is

$$\begin{aligned} J(\rho_\gamma) &= E_{\mathrm{H}}(\rho_\gamma) + E_{\mathrm{ext}}(\rho_\gamma) + E_{\mathrm{ZZ}} \\ &= \sup_{\phi \in \mathcal{V}} \left\{ \mathrm{Tr}(\Phi\gamma) - C_S \int_\Omega |\nabla \phi(\mathbf{r})|^2\,\mathrm{d}\mathbf{r} + \int_\Omega b(\mathbf{r},\{\mathbf{R}\})\phi(\mathbf{r})\,\mathrm{d}\mathbf{r} \right\} + C_{\mathrm{self}}. \end{aligned} \tag{12}$$

2.2.2 Exchange-correlation energy

Next we reformulate the exchange-correlation energy E_{xc}. We make the following assumptions on the integrand $h(t)$ in the exchange-correlation energy introduced in equation (10):

(P1) *Smoothness condition*: the function $h : \mathbb{R}_+ \to \mathbb{R}$ and $h(t) \in C^1(\mathbb{R}^3)$.

(P2) *Curvature condition*: the function h is concave in $\mathbb{R}^+$.

(P3) *Zero density condition*:

$$h(0) = 0. \tag{13}$$

(P4) *Non-positivity condition*: it holds $h(t) \leq 0$ for all $t \in \mathbb{R}^+$.

(P5) *Decay condition*: for $t \in \mathbb{R}^+$ the function h satisfies

$$h'(t) \leq 0. \tag{14}$$

(P6) *Growth conditions*: for $t \in \mathbb{R}^+$, the function h satisfies the bounds

$$C_1|t|^{4/3} + C_2 \leq |h(t)| \leq C_3|t|^{4/3} + C_4, \tag{15}$$

for some real constants $C_1 > 0$, $C_2 \leq 0$, $C_3 > 0$ and $C_4 \geq 0$.

By reflection, we can extend h to a function from $\mathbb{R}_+$ to $\mathbb{R}$, setting $h(t) \equiv h(|t|)$ for $t < 0$. This extended function, again denoted by h, is continuous in $\mathbb{R}$ due to property (P3).

Remark 2.2 *Since $h(t)$ is continuous in $\mathbb{R}$ and since $|h(t)| \leq C_3|t|^{\frac{4}{3}} + C_4$ from the upper bound in (15), with Fatou's Lemma it follows that $E_{\mathrm{xc}}(\rho_\gamma)$ is continuous in $\mathcal{L}^{\frac{4}{3}}(\mathbb{R}^3)$.*

We proceed to rewrite the exchange-correlation functional using a Legendre transform. We define

$$B_{\mathrm{xc}}(\rho_\gamma) = -E_{\mathrm{xc}}(\rho_\gamma).$$

From property (P2) of the exchange-correlation function h, $B_{\mathrm{xc}}(\rho_\gamma)$ is a convex and continuous functional in $\mathcal{L}^{4/3}(\mathbb{R}^3)$. Let

$$\mathcal{U} = \mathcal{L}^4(\mathbb{R}^3). \tag{16}$$

As explained in Appendix B, there exists a dual functional $B^*_{\mathrm{xc}}(u) : \mathcal{U} \mapsto \mathbb{R}$ such that

$$B_{\mathrm{xc}}(\rho_\gamma) = \sup_{u \in \mathcal{U}} \{\langle \rho_\gamma, u\rangle - B^*_{\mathrm{xc}}(u)\},$$

where the dual product $\langle v, u\rangle$ for any $v \in \mathcal{L}^{4/3}(\mathbb{R}^3)$ and $u \in \mathcal{L}^4(\mathbb{R}^3)$ is defined by

$$\langle v, u\rangle = \int_\Omega v(\mathbf{r})u(\mathbf{r})\,\mathrm{d}\mathbf{r}.$$

Using arguments from [13], we can rewrite the exchange-correlation functional,

$$\begin{aligned} E_{\mathrm{xc}}(\rho_\gamma) &= -B_{\mathrm{xc}}(\rho_\gamma) \\ &= -\sup_{u \in \mathcal{U}} \{\langle \rho_\gamma, u\rangle - B^*_{\mathrm{xc}}(u)\} \\ &= \inf_{u \in \mathcal{U}} \{-\langle \rho_\gamma, u\rangle + B^*_{\mathrm{xc}}(u)\}. \end{aligned}$$

Finally, we introduce the unbounded local operator

$$U(\mathbf{r},\mathbf{r}')=u(\mathbf{r})\delta(\mathbf{r},\mathbf{r}'), \tag{17}$$

using its coordinate representation. We can then rewrite the exchange-correlation functional as

$$E_{\text{xc}}(\rho_\gamma)=\inf_{u\in\mathcal{U}}\{-\text{Tr}(U\gamma)+B^*_{\text{xc}}(u)\}. \tag{18}$$

2.2.3 Reformulated Extended Kohn-Sham Functional

Substituting (12) and (18) in (6) and omitting the inessential constant C_{self} for brevity, we obtain the reformulated extended K-S(REKS) energy functional $E^{\text{REKS}}:\mathcal{K}_N\to\mathbb{R}$ as

$$E^{\text{REKS}}(\gamma)=\inf_{u\in\mathcal{U}}\sup_{\phi\in\mathcal{V}}L(u,\phi,\gamma), \tag{19}$$

where $L:\mathcal{U}\times\mathcal{V}\times\mathcal{K}_N$ is

$$L(u,\phi,\gamma)=\text{Tr}(H(\phi,u)\gamma)+\int_\Omega\big(-C_S|\nabla\phi(\mathbf{r})|^2+b(\mathbf{R},\mathbf{r})\phi(\mathbf{r})\big)\,\text{d}\mathbf{r}+B^*_{\text{xc}}(u), \tag{20}$$

with the Hamiltonian

$$H(\phi,u)=-\frac{1}{2}\Delta+\Phi-U$$

and Φ, U defined in (11), (17).

The ground-state energy of the system with M atoms is

$$\begin{aligned}\varepsilon^{\text{REKS}}_{\text{GS}}&=\inf_{\gamma\in\mathcal{K}_N}E^{\text{REKS}}(\gamma)\\&=\inf_{\gamma\in\mathcal{K}_N}\inf_{u\in\mathcal{U}}\sup_{\phi\in\mathcal{V}}L(u,\phi,\gamma)\\&=\inf_{\gamma\in\mathcal{K}_N}\inf_{u\in\mathcal{U}}\sup_{\phi\in\mathcal{V}}\left\{\text{Tr}(H(\phi,u)\gamma)+\int_\Omega\big(-C_S|\nabla\phi(\mathbf{r})|^2+b(\mathbf{r},\{\mathbf{R}\})\phi(\mathbf{r})\big)\,\text{d}\mathbf{r}+B^*_{\text{xc}}(u)\right\}.\end{aligned} \tag{21}$$

3 Main results

We have the following theorems on the reformulated extended K-S functional.

Theorem 1 *The reformulated extended K-S energy functional $E^{REKS}(\gamma)$ in (19) possesses a minimizer in $\mathcal{K}_N$.*

Theorem 2 *The order of the infimum and supremum in the computation of the ground-state energy of the reformulated K-S energy functional (21) can be exchanged,*

$$\begin{aligned}\varepsilon^{\text{REKS}}_{\text{GS}}&=\inf_{\gamma\in\mathcal{K}_N}\inf_{u\in\mathcal{U}}\sup_{\phi\in\mathcal{V}}L(u,\phi,\gamma)\\&=\inf_{u\in\mathcal{U}}\sup_{\phi\in\mathcal{V}}\inf_{\gamma\in\mathcal{K}_N}L(u,\phi,\gamma),\end{aligned} \tag{22}$$

where L is given by (20).

Theorem 2 enables the spectral discretization. Note that γ appears linearly in the functional L and only in $\mathrm{Tr}(H(\phi,u)\gamma)$. It is easy to show that for every $u \in \mathcal{U}$ and every $\phi \in \mathcal{V}$,

$$\inf_{\gamma \in \mathcal{K}_N} \mathrm{Tr}(H(\phi,u)\gamma)$$

is attained and the minimizer commutes with γ. Therefore the problem is unchanged if we seek the infimum over a subset $\mathcal{K}_N^H \subset \mathcal{K}_N$ of operators that commute with H or equivalently over the Borel functions of H (see (41) below). We obtain a spectral discretization by limiting γ to $\mathcal{K}_{N,k}^H$ made of k simple functions of H (see (49) below).

We are also interested in spatial discretization. Hence we consider finite-dimensional subspaces $\mathcal{V}_j$ and $\mathcal{U}_j$ of $\mathcal{V}$ and $\mathcal{U}$ respectively, with H^j, L^j to be discrete Hamiltonian and functional on these subspaces.

We have the following result on the combined convergence with respect to spatial and spectral discretization.

Theorem 3 *Let $k_j \to \infty$ as $j \to \infty$. Then, the diagonal sequence of spatially and spectrally discrete reformulated extended K-S energies converges to the full K-S ground-state energy,*

$$\lim_{j\to\infty} \inf_{\mathcal{U}_j} \sup_{\mathcal{V}_j} \inf_{\mathcal{K}_{N,k_j}^{H^j(\phi,u)}} L^j(u,\phi,\gamma) = \inf_{\mathcal{U}} \sup_{\mathcal{V}} \inf_{\mathcal{K}_N^{H(\phi,u)}} L(u,\phi,\gamma) = \varepsilon_{\mathrm{GS}}^{\mathrm{REKS}}.$$

4 Existence of solutions

To establish the existence of minimizers in $\mathcal{K}_N$ for the KS-DFT problem in equation (19), we use similar tools as the more general proof given by Anantharaman and Cancès in [2] and restate their results for an open, bounded, and Lipschitz domain Ω for completeness.

The proof follows the framework of the direct method in the calculus of variations. We consider the weak*-topology of the vector space $\mathcal{X}$ endowed with the norm

$$\|\cdot\|_{\mathcal{X}} = \mathrm{Tr}(|\cdot|) + \mathrm{Tr}(|\nabla|\cdot|\nabla||)$$

in the convex set $\mathcal{K}_N$ defined in (4).

For the clarity of notation, in the remainder of this paper, we change our notation on the repulsive energy functionals (10), (12) as to emphasize their dependence on the reduced one-particle density operator,

$$\begin{aligned} E_{\mathrm{xc}}(\gamma) &\equiv E_{\mathrm{xc}}(\rho_\gamma), \\ J(\gamma) &\equiv J(\rho_\gamma). \end{aligned}$$

Remark 4.1 *Since $\mathcal{X}$ is a separable and normed linear space, every uniformly bounded sequence $\{\gamma_n\}_{n\in\mathbb{N}}$ in $\mathcal{X}$ contains a weak*-convergent subsequence.*

For a proof of Remark 4.1, see for instance Part II of Theorem 2.2.1 in [15].

Let vol(Ω) denote the 3-dimensional Lebesgue measure of the bounded domain Ω.

Lemma 4.2 *For all $\gamma \in \mathcal{K}_N$, the following inequalities hold.*

1. Lower bound on the kinetic energy,

$$\frac{1}{2}\|\nabla\sqrt{\rho_\gamma}\| \leq \mathrm{Tr}(-\frac{1}{2}\Delta\gamma) = \frac{1}{2}\mathrm{Tr}(|\nabla|\gamma|\nabla|). \tag{23}$$

2. Lower bound on the Coulomb energy,

$$0 \leq J(\gamma).$$

3. Lower bound on the exchange-correlation energy,

$$-C_3(\mathrm{vol}\Omega)^{-1/3}N^{4/3} - C_4(\mathrm{vol}\Omega) \leq E_{\mathrm{xc}}(\gamma). \tag{24}$$

4. Lower bound on the reformulated extended K-S energy functional,

$$\|\gamma\|_{\mathcal{X}} - C_5 \leq E^{\mathrm{REKS}}(\gamma) \tag{25}$$

for a constant $C_5 > 0$ independent of γ. In particular, by (25), $E^{\mathrm{REKS}}(\gamma)$ is coercive w.r.t. the weak-topology of $\mathcal{X}$.*

Proof 1. *Lower bound on the kinetic energy.* In the canonical representation, the electron density is

$$\rho_\gamma(\mathbf{r}) = \sum_{i=1}^{\infty} 2\alpha_i\xi_i(\mathbf{r})^2.$$

By direct inspection and Cauchy–Schwarz's inequality, we find

$$\begin{aligned}|\nabla\sqrt{\rho_\gamma}|^2 &= \frac{2|\sum_{i=1}^{\infty}\alpha_i\xi_i(\mathbf{r})\nabla\xi_i(\mathbf{r})|^2}{\sum_{i=1}^{\infty}\alpha_i\xi_i(\mathbf{r})^2}\\ &\leq \frac{2\sum_{i=1}^{\infty}\alpha_i|\xi_i(\mathbf{r})|^2\sum_{i=1}^{\infty}\alpha_i|\nabla\xi_i(\mathbf{r})|^2}{\sum_{i=1}^{\infty}\alpha_i\xi_i(\mathbf{r})^2}.\end{aligned}$$

After integration, this yields

$$\frac{1}{2}\|\nabla\sqrt{\rho_\gamma}\|_{\mathcal{L}^2(\Omega)} \leq \mathrm{Tr}(-\frac{1}{2}\Delta\gamma) = \frac{1}{2}\mathrm{Tr}(|\nabla|\gamma|\nabla|).$$

2. *Lower bound on the Coulomb energy.* It holds

$$J(\gamma) = \sup_{\phi\in\mathcal{V}}\left\{\int_\Omega \phi(\mathbf{r})(b(\{\mathbf{R}\},\mathbf{r}) + \rho_\gamma(\mathbf{r}))\,\mathrm{d}\mathbf{r} - C_S\int_\Omega |\nabla\phi(\mathbf{r})|^2\,\mathrm{d}\mathbf{r}\right\} \geq 0, \tag{26}$$

where we use the test function $\phi(\mathbf{r}) = 0$ in Ω to obtain the lower bound.

3. *Lower bound on the exchange-correlation energy.*

   Using the bounds from equation (89) in Appendix B, the LDA exchange-correlation functional integrand h in equation (10) is bounded from below,

$$
\begin{aligned}
E_{\mathrm{xc}}(\gamma) &= \inf_{u\in\mathcal{U}}\{-\mathrm{Tr}(U\gamma) + B^*_{\mathrm{xc}}(u)\} \\
&\geq \inf_{u\in\mathcal{U}}\{-\mathrm{Tr}(U\gamma) + C_{18}\|u\|^4_{\mathcal{L}^4(\Omega)} + C_{19}(\mathrm{vol}\Omega)\} \\
&= -\mathrm{Tr}(U_\gamma\gamma) + C_{18}\|u_\gamma\|^4_{\mathcal{L}^4(\Omega)} + C_{19}(\mathrm{vol}\Omega) \qquad (27) \\
&\geq -\mathrm{Tr}(U_\gamma\gamma) + C_{19}(\mathrm{vol}\Omega) \\
&\geq -C(\mathrm{vol}\Omega)^{-1/3}\big(\mathrm{Tr}(\gamma)\big)^{4/3} + C_{19}(\mathrm{vol}\Omega) \\
&= -C(\mathrm{vol}\Omega)^{-1/3}N^{4/3} + C_{19}(\mathrm{vol}\Omega), \qquad (28)
\end{aligned}
$$

   where u_γ denotes a minimizer of equation (27) and U_γ is its corresponding operator. It is evident that there exists a minimizer for the variational problem (27).

4. *Lower bound on E^{REKS}. Coercivity of E^{REKS}.*

   Putting together all the inequalities in the equations (26) and (28), we end up with

$$
E^{\mathrm{REKS}}(\gamma) \geq \mathrm{Tr}\Big(-\frac{1}{2}\Delta\gamma\Big) - C(\mathrm{vol}\Omega)^{-1/3}N^{4/3} + C_{19}(\mathrm{vol}\Omega) = \frac{1}{2}\big(\mathrm{Tr}(|\nabla|\gamma|\nabla|) + \mathrm{Tr}(|\gamma|)\big) - C_5. \qquad (29)
$$

   Here, we introduced the new constant

$$
C_5 \equiv C(\mathrm{vol}\Omega)^{-1/3}N^{4/3} - C_{19}(\mathrm{vol}\Omega) + \frac{N}{2}.
$$

   For the derivation of (29), we used that for every $\gamma \in \mathcal{K}_N$, directly from the definition of this set,

$$
\mathrm{Tr}(\gamma) = \mathrm{Tr}(|\gamma|) = N.
$$

   The estimate (29) implies that for any $t \in \mathbb{R}$ the level sets

$$
\{\gamma \in \mathcal{K}_N : E^{\mathrm{REKS}}(\gamma) \leq t\}
$$

   are bounded,

$$
t + C_5 \geq \frac{1}{2}(\mathrm{Tr}(|\gamma|) + \mathrm{Tr}(|\nabla|\gamma|\nabla|)) \equiv \frac{1}{2}\|\gamma\|_{\mathcal{X}}.
$$

   Consequently there exists a subsequence of γ_n that converges w.r.t. the weak*-topology and we conclude that $E^{\mathrm{REKS}}(\gamma)$ is coercive w.r.t. the weak*-topology in $\mathcal{K}_N$. □

Lemma 4.3 *The set $\mathcal{K}_N$ is closed in $\mathcal{X}$ w.r.t. the weak*-topology.*

Proof Let $\mathfrak{C}(\mathcal{H})$ denote the vector space of compact linear operators on $\mathcal{H}$. For all $\gamma_n \overset{*}{\rightharpoonup} \gamma$, we have $\mathrm{Tr}(\gamma_n W) \to \mathrm{Tr}(\gamma W)$ for all $W \in \mathfrak{C}(\mathcal{H})$ in the limit $n \to \infty$.

We define the rank-one operator

$$
W = |\psi\rangle\langle\psi|,
$$

where $\|\psi\|_{\mathcal{L}^2(\Omega)} = 1$. Due to the weak*-convergence of γ_n,

$$
0 \leq \lim_{n\to\infty} \mathrm{Tr}(\gamma_n W) = \mathrm{Tr}(\gamma W), \qquad (30)
$$

and

$$\mathrm{Tr}(\gamma W) = \lim_{n\to\infty} \mathrm{Tr}(\gamma_n W) = \lim_{n\to\infty} \langle \psi, \gamma_n \psi\rangle \leq \langle \psi, \psi \rangle = 1. \tag{31}$$

Since the estimate (31) holds for all normalized $\psi \in \mathcal{H}$, we find with (30) that $0 \leq \gamma \leq 1$.

Since $\gamma_n \overset{*}{\rightharpoonup} \gamma$, $\|\gamma_n\|_1$ is bounded independently of n, see Proposition 3.13 in [9]. From equation (23) we have that $\{\sqrt{\rho_{\gamma_n}}\}_{n\in\mathbb{N}}$ is bounded in $\mathcal{W}_0^{1,2}(\Omega)$. Therefore there exists a subsequence $\{\sqrt{\rho_{\gamma_{n_i}}}\}_{i\in\mathbb{N}}$ that converges weakly to $\sqrt{\rho_\gamma}$ in $\mathcal{W}_0^{1,2}(\Omega)$. By the compact embedding of $\mathcal{W}_0^{1,2}(\Omega)$ in $\mathcal{L}^p(\Omega)$, the subsequence $\{\sqrt{\rho_{\gamma_{n_i}}}\}_{i\in\mathbb{N}}$ converges strongly to $\sqrt{\rho_\gamma}$ in $\mathcal{L}^p(\Omega)$ for all $2 \leq p < 6$, see, e.g., [1]. These considerations show that

$$\lim_{n\to\infty} \mathrm{Tr}(\gamma_n) = \lim_{n\to\infty} \int_\Omega \rho_{\gamma_n}\,\mathrm{d}\mathbf{r} = \lim_{n\to\infty} \|\sqrt{\rho_{\gamma_n}}\|^2_{\mathcal{L}^2} = \|\sqrt{\rho_\gamma}\|^2_{\mathcal{L}^2(\Omega)} = \int_\Omega \rho_\gamma\,\mathrm{d}\mathbf{r} = \mathrm{Tr}(\gamma).$$

Hence the set $\mathcal{K}_N$ is closed w.r.t. the weak*-topology on $\mathcal{X}$. □

Lemma 4.4 *The functional $J(\gamma)$ introduced in (12) is lower semi-continuous w.r.t. the weak*-topology on $\mathcal{X}$.*

Proof We begin by showing that $\mathrm{Tr}(\Phi\cdot)$ defines a bounded linear functional on $\mathcal{K}_N$,

$$\begin{aligned} |\mathrm{Tr}(\Phi\gamma)| &= \Big|\sum_{i=1}^\infty \langle \Phi\gamma\xi_i, \xi_i\rangle\Big| \leq \sum_{i=1}^\infty 2\alpha_i |\langle \Phi\xi_i, \xi_i\rangle| \\ &\leq \sum_{i=1}^\infty 2\alpha_i \|\phi\|_{\mathcal{L}^2(\Omega)} \|\xi_i^2\|_{\mathcal{L}^2(\Omega)} = \|\phi\|_{\mathcal{L}^2(\Omega)} \sum_{i=1}^\infty 2\alpha_i \|\xi_i\|^2_{\mathcal{L}^4(\Omega)} \\ &\leq C\|\phi\|_{\mathcal{L}^2(\Omega)} \sum_{i=1}^\infty 2\alpha_i \|\nabla\xi_i\|^2_{\mathcal{L}^2(\Omega)} = C\|\phi\|_{\mathcal{L}^2(\Omega)} \mathrm{Tr}(-\Delta\gamma), \end{aligned} \tag{32}$$

where $\{\xi_i\}_{i\in\mathbb{N}}$ come from the canonical representation of $\gamma \in \mathcal{K}_N$, cf. equation (5), and the Gagliardo–Nirenberg–Sobolev inequality has been used to obtain equation (32). Consequently,

$$J(\gamma) = \sup_{\phi\in\mathcal{V}} \left\{ \mathrm{Tr}(\Phi\gamma) + \int_\Omega \big(b(\mathbf{r}, \{\mathbf{R}\})\phi(\mathbf{r}) - C_S |\nabla\phi(\mathbf{r})|^2\big) d\mathbf{r} \right\}$$

is the point-wise supremum over a family of continuous affine functionals on $\mathcal{K}_N$. Hence it is also lower semi-continuous with respect to the weak*-topology on $\mathcal{K}_N$. □

Lemma 4.5 *$E_{\mathrm{xc}}(\gamma)$ is continuous w.r.t. the weak*-topology on $\mathcal{X}$.*

Proof Similar to the proof in Lemma 4.4, we can show that $\mathrm{Tr}(U\gamma)$ defines a continuous affine functional on $\mathcal{K}_N$ for every $u \in \mathcal{U}$. We prove the continuity of $E_{\mathrm{xc}}(\gamma)$ with respect to the weak*-topology using techniques of Γ-convergence.

For every sequence γ_n such that $\gamma_n \overset{*}{\rightharpoonup} \gamma$ in $\mathcal{K}_N$, we consider the family of functionals on $\mathcal{U}$ indexed by n defined by

$$-\mathrm{Tr}(U\gamma_n) + B^*_{\mathrm{xc}}(u).$$

We show that this family of functionals Γ-converges with respect to the weak*-topology to the functional

$$-\mathrm{Tr}(U\gamma) + B^*_{\mathrm{xc}}(u)$$

for all $\gamma_n \overset{*}{\rightharpoonup} \gamma$ in $\mathcal{K}_N$.

For the lim-inf condition, we need to show that for every $u \in \mathcal{U}$ and for all $u_n \rightharpoonup u$,

$$\liminf_{n\to\infty}\{-\mathrm{Tr}(U_n\gamma_n) + B^*_{\mathrm{xc}}(u_n)\} \geq -\mathrm{Tr}(U\gamma) + B^*_{\mathrm{xc}}(u).$$

Since $\gamma_n \overset{*}{\rightharpoonup} \gamma$, for every member of a complete orthonormal basis in $\mathcal{L}^2(\Omega)$, $\{\xi_i\}_{i\in\mathbb{N}} \subset \mathcal{W}^{1,2}_0(\Omega)$, we have

$$\lim_{n\to\infty}\langle\gamma_n\xi_i, v\rangle = \langle\gamma\xi_i, v\rangle.$$

From the proof of Lemma 4.3, we have $\rho_{\gamma_n} \to \rho_\gamma$ in $\mathcal{L}^2(\Omega)$. Therefore $\lim_{n\to\infty}\mathrm{Tr}(U_n\gamma_n) = \mathrm{Tr}(U\gamma)$. In addition, $B^*_{\mathrm{xc}}(u)$ is weakly lower semi-continuous by duality and convexity. So, the lim-inf condition is proven.

For the lim-sup condition, we choose the trivial recovery sequence $u_n = u$ for every $u \in \mathcal{U}$, implying

$$\limsup_{n\to\infty}\{-\mathrm{Tr}(U_n\gamma_n) + B^*_{\mathrm{xc}}(u_n)\} \geq -\mathrm{Tr}(U\gamma) + B^*_{\mathrm{xc}}(u).$$

Lastly, to show equi-coercivity of the functionals, from equation (89) in Appendix B,

$$-\mathrm{Tr}(u\gamma_n) + B^*_{\mathrm{xc}}(u) \geq C_{18}\|u\|^4_{\mathcal{U}} - (\sup_n C_n)\|u\|_{\mathcal{L}^2(\Omega)} + C_{19}(\mathrm{vol}\Omega),$$

where $C_n \equiv \mathrm{Tr}(-\Delta\gamma_n)$, and C_n is bounded since $\gamma_n \overset{*}{\rightharpoonup} \gamma$ in $\mathcal{X}$. Therefore the family of functionals

$$-\mathrm{Tr}(u\gamma_n) + B^*_{\mathrm{xc}}(u)$$

is equi-coercive. Using Theorem 7.8 in [12], we have

$$\lim_{n\to\infty} E_{\mathrm{xc}}(\gamma_n) = \lim_{n\to\infty}\inf_{u\in\mathcal{U}}\{-\mathrm{Tr}(U\gamma_n) + B^*_{\mathrm{xc}}(u)\} = \inf_{u\in\mathcal{U}}\{\mathrm{Tr}(U\gamma) + B^*_{\mathrm{xc}}(u)\} = E_{\mathrm{xc}}(\gamma). \qquad \square$$

Lemma 4.6 *Let $\{\gamma_n\}_{n\in\mathbb{N}}$ be a sequence of elements in $\mathcal{K}_N$ which converges to γ in the weak*-topology of $\mathcal{X}$. Then*

$$E^{\mathrm{REKS}}(\gamma) \leq \liminf_{n\to\infty} E^{\mathrm{REKS}}(\gamma_n).$$

Proof To prove the lower semi-continuity of $E^{\mathrm{REKS}}(\gamma)$, we use the continuity of the functional $J(\gamma)$ from Lemma 4.4 and the continuity of $E_{\mathrm{xc}}(\gamma)$ from Remark 2.2 w.r.t. the weak*-topology.

For any orthonormal basis $\{\psi_k\}_{k\in\mathbb{N}}$ of $\mathcal{L}^2(\Omega)$ such that $\psi_k \in \mathcal{W}^{1,2}(\Omega)$ for all k, we have

$$\begin{aligned}
\mathrm{Tr}(-\Delta\gamma) &= \mathrm{Tr}(|\nabla|\gamma|\nabla|) \\
&= \sum_{k=1}^{\infty}\langle\psi_k\big||\nabla|\gamma|\nabla|\big|\psi_k\rangle \\
&= \sum_{k=1}^{\infty}\mathrm{Tr}\big(\gamma(\big||\nabla|\psi_k\rangle\langle|\nabla|\psi_k\big|)\big) \\
&= \sum_{k=1}^{\infty}\lim_{n\to\infty}\mathrm{Tr}\big(\gamma_n(\big||\nabla|\psi_k\rangle\langle|\nabla|\psi_k\big|)\big) \\
&\leq \liminf_{n\to\infty}\sum_{k=1}^{\infty}\mathrm{Tr}\big(\gamma_n(\big||\nabla|\psi_k\rangle\langle|\nabla|\psi_k\big|)\big) \\
&= \liminf_{n\to\infty}\mathrm{Tr}(|\nabla|\gamma_n|\nabla|).
\end{aligned} \tag{33}$$

This proves the lower semi-continuity of the functional $E^{\mathrm{REKS}}(\gamma)$. $\square$

Theorem 1 *The reformulated extended K-S energy functional $E^{\text{REKS}}(\gamma)$ possesses a minimizer in $\mathcal{K}_N$.*

Proof Consider a minimizing sequence $\{\gamma_n\}_{n\in\mathbb{N}}$ of $E^{\text{REKS}}(\gamma)$ in $\mathcal{K}_N$. From Lemma 4.2 and Lemma 4.1, we know that $(\gamma_n)_{n\in\mathbb{N}}$ has a weak*-converging subsequence. By the closure of the subset $\mathcal{K}_N$, this subsequence converges to some $\gamma_0 \in \mathcal{K}_N$. Using the lower semi-continuity of E^{REKS} w.r.t. the weak*-convergence in $\mathcal{X}$, it follows

$$\inf_{\gamma\in\mathcal{K}_N} E^{\text{REKS}}(\gamma) \leq E^{\text{REKS}}(\gamma_0) \leq \liminf_{n\to\infty} E^{\text{REKS}}(\gamma_n) = \inf_{\gamma\in\mathcal{K}_N} E^{\text{REKS}}(\gamma).$$

Hence the existence of a minimizer of E^{REKS} in $\mathcal{K}_N$ is established. □

5 Discretization of the energy functional

Next we introduce both the spectral and spatial discretization of the reformulated extended K-S energy functional and prove the convergence of simultaneously discretizing the energy functional both spatially and spectrally.

5.1 Justification of the spectral discretization

Before we can apply spectral discretization, as it will be evident subsequently, we have to prove that the spinless one-particle density operator that minimizes $E^{REKS}(\gamma)$ can be written as a spectral function of the Hamiltonian $H(\phi, u)$.

We recall the definition of $L : \mathcal{U} \times \mathcal{V} \times \mathcal{K}_N$ from equation (20),

$$L(u, \phi, \gamma) = \text{Tr}(H(\phi, u)\gamma) + \int_\Omega \big(- C_S|\nabla\phi(\mathbf{r})|^2 + b(\{\mathbf{R}\}, \mathbf{r})\phi(\mathbf{r})\big)\, \mathrm{d}\mathbf{r} + B^*_{\text{xc}}(u).$$

The ground-state energy equals, cf. the equations (19) and (20),

$$\varepsilon^{\text{REKS}}_{\text{GS}} = \inf_{\gamma\in\mathcal{K}_N} \inf_{u\in\mathcal{U}} \sup_{\phi\in\mathcal{V}} L(u, \phi, \gamma).$$

Since we can exchange the order of the infima, the ground-state energy is also equal to

$$\varepsilon^{\text{REKS}}_{\text{GS}} = \inf_{u\in\mathcal{U}} \inf_{\gamma\in\mathcal{K}_N} \sup_{\phi\in\mathcal{V}} L(u, \phi, \gamma). \tag{34}$$

Now we derive sufficient properties of $L(u, \cdot, \cdot)$ that enable us to exchange the order of the infimum over $\gamma \in \mathcal{K}_N$ and the supremum over $\phi \in \mathcal{V}$.

Lemma 5.1 *For every $u \in \mathcal{U}$ and every $\phi \in \mathcal{V}$, the functional $L(u, \phi, \cdot)$ is convex and lower semi-continuous with respect to γ in $\mathcal{X}$. In addition, for every $\phi \in \mathcal{V}$,*

$$\lim_{\|\gamma\|_{\mathcal{X}}\to+\infty} L(u, \phi, \gamma) = +\infty. \tag{35}$$

Proof For given u and ϕ, the convexity of $L(u, \phi, \cdot)$ is evident since the terms involving γ are linear functionals of γ.

Regarding the lower semi-continuity of $L(u, \phi, \cdot)$, from Lemma 4.6 we observe that $\text{Tr}(-\frac{1}{2}\Delta\gamma)$ is lower semi-continuous in $\mathcal{X}$. Since for every sequence $\gamma_n \to \gamma$ in $\mathcal{K}_N$ by compact embedding it holds $\rho_{\gamma_n} \to \rho_\gamma$ in $\mathcal{L}^2(\Omega)$, the functionals $\text{Tr}(\Phi\gamma)$ and $\text{Tr}(U\gamma)$ are also continuous in $\mathcal{X}$.

Since $u \in \mathcal{U} \subset \mathcal{L}^2(\Omega)$, it holds for every $\gamma \in \mathcal{K}_N$,

$$\begin{aligned}
L(u, \phi, \gamma) &= \mathrm{Tr}(-\frac{1}{2}\Delta\gamma) + \mathrm{Tr}(\Phi\gamma) - \mathrm{Tr}(U\gamma) \\
&\geq \mathrm{Tr}(-\frac{1}{2}\Delta\gamma) - (\|u\|_{\mathcal{L}^2(\Omega)} + \|\phi\|_{\mathcal{L}^2(\Omega)})\|\rho_\gamma\|_{\mathcal{L}^2(\Omega)} \\
&\geq \mathrm{Tr}(-\frac{1}{2}\Delta\gamma) - C_6(\|u\|_{\mathcal{L}^2(\Omega)} + \|\phi\|_{\mathcal{L}^2(\Omega)})\|\rho_\gamma\|^{\frac{1}{4}}_{\mathcal{L}^1(\Omega)}\|\rho_\gamma\|^{\frac{3}{4}}_{\mathcal{L}^3(\Omega)} \qquad (36) \\
&\geq \mathrm{Tr}(-\frac{1}{2}\Delta\gamma) - C_7(\|u\|_{\mathcal{L}^2(\Omega)} + \|\phi\|_{\mathcal{L}^2(\Omega)})\mathrm{Tr}(|\gamma|)^{\frac{1}{4}}\|\nabla\sqrt{\rho_\gamma}\|^{\frac{3}{2}}_{\mathcal{L}^2(\Omega)} \qquad (37)
\end{aligned}$$

for some positive real constants C_6 and C_7, where interpolation inequalities are used to obtain equation (36) and the Gagliardo–Nirenberg–Sobolev inequality is used to obtain equation (37). Hence

$$L(u, \phi, \gamma) \geq \frac{1}{2}\|\gamma\|_{\mathcal{X}} - C_8\|\nabla\sqrt{\rho_\gamma}\|^{\frac{3}{2}}_{\mathcal{L}^2(\Omega)} - \frac{N}{2},$$

where $C_8 \equiv C_7 N^{1/4}(\|u\|_{\mathcal{L}^2(\Omega)} + \|\phi\|_{\mathcal{L}^2(\Omega)})$, implying the coercivity (35) of $L(u, \phi, \cdot)$. □

Lemma 5.2 *For every $u \in \mathcal{U}$ and every $\gamma \in \mathcal{K}_N$, the functional $L(u, \cdot, \gamma)$ is concave and upper semi-continuous with respect to ϕ in $\mathcal{V}$. In addition,*

$$\lim_{\|\phi\|_{\mathcal{V}} \to +\infty} L(u, \phi, \gamma) = -\infty.$$

Proof For given u and γ, the terms $\mathrm{Tr}(\Phi\gamma)$ and $\int_\Omega b(\mathbf{r}, \{\mathbf{R}\})\phi(\mathbf{r})\,\mathrm{d}\mathbf{r}$ are linear functionals of ϕ, so they are concave. The term $-C_S\int_\Omega |\nabla\phi(\mathbf{r})|^2\,\mathrm{d}\mathbf{r}$ is quadratic and concave in $|\nabla\phi(\mathbf{r})|$. Hence, $L(u, \cdot, \gamma)$ is concave.

Concerning the upper semi-continuity of $L(u, \cdot, \gamma)$, by using arguments similar to those in Lemma 5.1, we observe that $\mathrm{Tr}(\Phi\gamma)$ and $\int_\Omega b(\mathbf{r}, \{\mathbf{R}\})\phi(\mathbf{r})\,\mathrm{d}\mathbf{r}$ are continuous in $\mathcal{V}$ for given $b(\mathbf{r}, \{\mathbf{R}\})$ and $\gamma \in \mathcal{K}_N$. The quadratic term $-C_S\int_\Omega |\nabla\phi(\mathbf{r})|^2\,\mathrm{d}\mathbf{r}$ is upper semi-continuous in $\mathcal{V}$ as a result of Proposition 2.1 in [12].

Finally, for every $\gamma \in \mathcal{K}_N$,

$$\begin{aligned}
-L(u, \phi, \gamma) &\geq C_S\|\nabla\phi\|^2_{\mathcal{L}^2(\Omega)} - \|\phi\|_{\mathcal{L}^2(\Omega)}\|\rho_\gamma + b(\mathbf{r}, \{\mathbf{R}\})\|_{\mathcal{L}^2(\Omega)} + C_9(u, \gamma) \\
&\geq C_{10}\|\phi\|^2_{\mathcal{L}^2(\Omega)} - \|\phi\|_{\mathcal{L}^2(\Omega)}\|\rho_\gamma + b(\mathbf{r}, \{\mathbf{R}\})\|_{\mathcal{L}^2(\Omega)} + C_9(u, \gamma), \qquad (38)
\end{aligned}$$

where the Poincaré inequality has been used to derive the second estimate, $C_{10} > 0$, and with

$$C_9(u, \gamma) \equiv \mathrm{Tr}(\frac{1}{2}\Delta\gamma) + \mathrm{Tr}(U\gamma) - B^*_{\mathrm{xc}}(u).$$

Applying Young's inequality to $\|\phi\|_{\mathcal{L}^2(\Omega)}\|\rho_\gamma + b(\mathbf{r}, \{\mathbf{R}\})\|_{\mathcal{L}^2(\Omega)}$ in (38), $\|\phi\|_{\mathcal{L}^2(\Omega)}$ can be absorbed in $C_{10}\|\phi\|^2_{\mathcal{L}(\Omega)}$, implying the convergence of $\phi \mapsto L(u, \phi, \gamma)$ to $-\infty$ as $\|\phi\|_{\mathcal{V}}$ converges to $+\infty$. □

After these ancillary results, we can now show the second main theorem which deals with exchanging the orders of infima and supremum when computing $\varepsilon^{\mathrm{REKS}}_{\mathrm{GS}}$. Theorem 2 is important as it allows to apply spectral theory to the Lagrange functional $L(u, \phi, \gamma)$.

Let $E_{\text{band}}(u,\phi,\gamma) := \text{Tr}(H(u,\phi)\gamma)$.

Theorem 2 *The order of the infimum and supremum in the computation of the ground-state energy of the reformulated K-S energy functional can be exchanged,*

$$\begin{aligned}\varepsilon_{\text{GS}}^{\text{REKS}} &= \inf_{\gamma\in\mathcal{K}_N} \inf_{u\in\mathcal{U}} \sup_{\phi\in\mathcal{V}} L(u,\phi,\gamma)\\ &= \inf_{u\in\mathcal{U}} \sup_{\phi\in\mathcal{V}} \inf_{\gamma\in\mathcal{K}_N} L(u,\phi,\gamma)\\ &= \inf_{u\in\mathcal{U}} \sup_{\phi\in\mathcal{V}} \inf_{\gamma\in\mathcal{K}_N} \left\{ E_{\text{band}}(u,\phi,\gamma) + \int_\Omega \big(- C_S|\nabla\phi(\mathbf{r})|^2 + b(\mathbf{R},\mathbf{r})\phi(\mathbf{r})\big)\,\mathrm{d}\mathbf{r} + B_{\text{xc}}^*(u)\right\}. \qquad (39)\end{aligned}$$

For every $u \in \mathcal{U}$ and every $\phi \in \mathcal{V}$, the minimizer of the band energy $E_{\text{band}}(u,\phi,\cdot)$ in $\mathcal{K}_N$ commutes with the Hamiltonian $H(\phi,u)$.

Proof Using similar arguments as in Proposition 2.2 in [13], we are guaranteed the existence of at least one saddle point $\{\bar{\phi},\bar{\gamma}\}$ of $L(u,\cdot,\cdot)$ for every $u \in \mathcal{U}$. Hence, exchanging infimum and supremum does not affect the ground-state energy of the reformulated K-S energy functional.

Next, for every $u \in \mathcal{U}$ and every $\phi \in \mathcal{V}$, $H(\phi,u)$ is a self-adjoint unbounded operator on $\mathcal{L}^2(\Omega)$. Associated to $H(\phi,u)$, there is a countable family of orthonormal eigenvectors that form a basis of $\mathcal{L}^2(\Omega)$. From [33], since $\phi(r) \in \mathcal{V}$ and $u(r) \in \mathcal{U}$, we have that $H(\phi,u)$ is semi-bounded from below.

Let λ_k, ξ_k denote the k-th eigenvalue and k-th eigenvector of $H(\phi,u)$, respectively, with the indices ordered by increasing magnitude of the eigenvalues. Then, since the trace is invariant with respect to a change of basis, it follows

$$\begin{aligned}\inf_{\gamma\in\mathcal{K}_N} E_{\text{band}}(u,\phi,\gamma) = \inf_{\gamma\in\mathcal{K}_N} \text{Tr}\big(H(\phi,u)\gamma\big) &= \inf_{\gamma\in K_N} \sum_{k=1}^{\infty} \langle H(\phi,u)\gamma\xi_k,\xi_k\rangle\\ &= \inf_{\gamma\in K_N} \sum_{k=1}^{\infty} \langle \gamma\xi_k, H(\phi,u)\xi_k\rangle\\ &= \inf_{\gamma\in K_N} \sum_{k=1}^{\infty} \lambda_k\langle \gamma\xi_k,\xi_k\rangle\\ &= \sum_{k=1}^{N} \lambda_k.\end{aligned}$$

From Theorem 1.3, Supplement 1 in [6], there exists a Borel function $g : \mathbb{R} \to \mathbb{R}$ with

$$g(\lambda) = \begin{cases} 1, & \text{if } \lambda \le \lambda_N,\\ 0, & \text{otherwise.}\end{cases}$$

such that for every $u \in \mathcal{U}$ and every $\phi \in \mathcal{V}$,

$$\operatorname*{argmin}_{\gamma\in\mathcal{K}_N} E_{\text{band}}(u,\phi,\gamma) = g\big(H(\phi,u)\big). \qquad (40)$$

To ensure the existence of a spectral function g, we replace the minimization over $\mathcal{K}_N$ by the minimization over the subset

$$\mathcal{K}_N^{H(\phi,u)} = \big\{\gamma \in \mathcal{K}_N : \gamma = g\big(H(\phi,u)\big) \text{ for a Borel function } g \text{ over } \mathbb{R},\ 0 \le g \le 1\big\} \qquad (41)$$

and observe that

$$\inf_{\gamma\in\mathcal{K}_N} E_{\text{band}}(u,\phi,\gamma) = \inf_{\gamma\in\mathcal{K}_N^{H(\phi,u)}} \text{Tr}\big(H(\phi,u)\gamma\big). \qquad \square \qquad (42)$$

We want to emphasize that every element in the set $\mathcal{K}_N^{H(\phi,u)}$ can be written as a spectral function of $H(\phi, u)$ and is thus amenable to spectral discretization.

We proceed in the next two sections to set up the spectral discretization and the spatial discretization of the reformulated extended K-S energy functional defined in (19).

5.2 Spatial discretization

We proceed to discretize problem (34) *à la* Rayleigh-Ritz, i.e. by restriction to finite-dimensional subspaces. To this end, let $\mathcal{V}_j$ be from a family of finite-dimensional subspaces of $\mathcal{V}$ spanned by the basis $\{e_1, \ldots, e_j\}$, e.g. a subspace that corresponds to a finite element discretization, and let $\mathcal{U}_j$ be from a family of finite-dimensional subspaces of $\mathcal{U}$ spanned by the basis $\{d_1, \ldots, d_j\}$, e.g. the piece-wise constant simple functions. Then the restriction of the electrostatic field to $\mathcal{V}_j$ is of the form

$$\phi_j(\mathbf{r}) = \sum_{a=1}^{j} \phi_a e_a(\mathbf{r}).$$

The nuclear charge distribution is

$$b_j(\mathbf{r}, \{\mathbf{R}\}) = \sum_{a=1}^{j} b_a^{\{\mathbf{R}\}} e_a(\mathbf{r})$$

and the dual density potential $u_j(\mathbf{r})$ has the form

$$u_j(\mathbf{r}) = \sum_{a=1}^{j} u_a d_a(\mathbf{r}).$$

Like-wise, the restricted density operator on a finite-dimensional subspace, the discrete density matrix, is

$$\gamma_j(\mathbf{r}_1, \mathbf{r}_2) = \sum_{a_1=1}^{j} \sum_{a_2=1}^{j} \gamma_{a_1,a_2}^{j} e_{a_1}(\mathbf{r}_1) e_{a_2}(\mathbf{r}_2), \tag{43}$$

where γ^j denotes the matrix of coefficients, and the discrete electron density follows as

$$\rho_j(\mathbf{r}) = \sum_{a_1=1}^{j} \sum_{a_2=1}^{j} \rho_{a_1 a_2}^{j} e_{a_1}(\mathbf{r}) e_{a_2}(\mathbf{r}),$$

where

$$\rho_{a_1 a_2}^{j} = \gamma_{a_1,a_2}^{j}.$$

The above restrictions define a sequence of subspaces in $\mathcal{K}_N^j$ of *density matrices*,

$$\mathcal{K}_N^j = \big\{\gamma \in \mathcal{X} : \gamma \in \mathcal{S}(\mathcal{V}_j),\ 0 \leq \gamma \leq 1\big\},$$

where $\mathcal{S}(\mathcal{V}_j)$ denotes the vector space of symmetric linear operators on $\mathcal{V}_j$.

The corresponding discrete Lagrangians L^j, obtained by restriction of the functional in equation (20) to $\mathcal{U}_j \times \mathcal{V}_j \times \mathcal{K}_N^j$, follow as

$$L^j(u, \phi, \gamma) = \mathrm{Tr}(H^j(\phi, u)\gamma^j) + \sum_{a_1=1}^{j} \sum_{a_2=1}^{j} \Big\{ - C_S \phi_{a_1} A_{a_1,a_2} \phi_{a_2} + b_{a_1}^{\{\mathbf{R}\}} \mathcal{M}_{a_1,a_2} \phi_{a_2} \Big\} + B_{\mathrm{xc}}^*(u). \tag{44}$$

Before proceeding further, we have to clarify our notation in (44). Let $H^j(\phi, u)$ denote the matrix H^j defined by restriction of ϕ and u on the finite-dimensional subspaces $\mathcal{V}_j$ and $\mathcal{U}_j$, respectively. Throughout this paper, we use a *superscript* index j to denote restriction of an operator or a functional to the finite-dimensional subspace defined by $\mathcal{V}_j$, $\mathcal{U}_j$ and $\mathcal{K}_N^j$. We use a *subscript* index j in general to denote the j-th element in a sequence of functions or operators. There will be cases where an operator or a function indexed by a *subscript* j happens to coincide with the restriction of the operator or the function to the finite-dimensional subspace $\mathcal{U}_j$,$\mathcal{V}_j$ and $\mathcal{K}_N^j$, but there is no ambiguity from the context when these situations arise.

Using spatial discretization, we introduce the discrete quantities,

$$
\begin{aligned}
H^j &\equiv \frac{1}{2}A + \Phi^j - U^j, \qquad (45)\\
A_{a_1,a_2} &\equiv \int_\Omega \nabla e_{a_1}(\mathbf{r}) \cdot \nabla e_{a_2}(\mathbf{r})\,\mathrm{d}\mathbf{r},\\
\mathcal{M}_{a_1,a_2} &\equiv \int_\Omega e_{a_1}(\mathbf{r}) \cdot e_{a_2}(\mathbf{r})\,\mathrm{d}\mathbf{r},\\
\Phi^j_{a_1,a_2} &\equiv \int_\Omega \Big(\sum_{a=1}^{j} \phi_a e_a(\mathbf{r})\Big) e_{a_1}(\mathbf{r}) e_{a_2}(\mathbf{r})\,\mathrm{d}\mathbf{r},\\
U^j_{a_1,a_2} &\equiv \int_\Omega \Big(\sum_{a=1}^{j} u_a d_a(\mathbf{r})\Big) e_{a_1}(\mathbf{r}) e_{a_2}(\mathbf{r})\,\mathrm{d}\mathbf{r}.
\end{aligned}
$$

Formally, A and $\mathcal{M}$ also depend on j as they are restrictions of operators to $\{e_1, \ldots, e_j\}$. We ignore this fact here to avoid heavy notation.

The discrete band energy $E^j_{\mathrm{band}} : \mathcal{U}_j \times \mathcal{V}_j \times \mathcal{K}_N^j$ becomes

$$
E^j_{\mathrm{band}}(u, \phi, \gamma) = \mathrm{Tr}(H^j(\phi, u)\gamma^j). \qquad (46)
$$

In addition, we need to introduce the sequence of discrete constraint sets,

$$
\mathcal{K}_N^{H^j(\phi,u)} = \big\{\gamma \in \mathcal{K}_N^j \,:\, \gamma = g\big(H^j(\phi, u)\big) \text{ for a Borel function } g \text{ over } \mathbb{R},\, 0 \le g \le 1\big\}.
$$

With these settings, motivated by the equations (19)–(21), the corresponding sequence of discrete energies $\varepsilon^{\mathrm{REKS}}_{\mathrm{GS},j}$ becomes

$$
\varepsilon^{\mathrm{REKS}}_{\mathrm{GS},j} = \inf_{u \in \mathcal{U}_j} \sup_{\phi \in \mathcal{V}_j} \inf_{\gamma \in \mathcal{K}_N^{H^j(\phi,u)}} L^j(u, \phi, \gamma). \qquad (47)
$$

5.3 Spectral discretization

Next we proceed to spectrally discretize the minimization over $\gamma \in \mathcal{K}_N^{H^j(\phi,u)}$ of the discrete band energy from equation (46). We begin by applying the spectral decomposition theorem (cf., e. g., [25]).

For fixed $j \in \mathbb{N}$, since H^j defined in (45) is a self-adjoint operator, this theorem states that

$$
H^j = \int_{\sigma(H^j)} \lambda\,\mathrm{d}P^j(\lambda),
$$

where P^j is a resolution of the identity over the Borel sets of the real line, and $\sigma(H^j)$ denotes the spectrum of H^j. Similarly, for the restricted discrete density matrices γ^j in (43) defined for H^j, there exist bounded Borel functions $g^j : \mathbb{R} \to \mathbb{R}$ with

$$\gamma^j = \int_{\sigma(H^j)} g^j(\lambda)\, \mathrm{d}P^j(\lambda).$$

Using this representation, we define

$$\begin{aligned} E^j(g^j) \equiv \mathrm{Tr}(H^j \gamma^j) &= \sum_{a=1}^{\infty} \int_{\sigma(H^j)} g^j(\lambda)\lambda\, \mathrm{d}\mu^j_{e_a,e_a}(\lambda), \\ N^j(g^j) \equiv \mathrm{Tr}(\gamma^j) &= \sum_{a=1}^{\infty} \int_{\sigma(H^j)} g^j(\lambda)\, \mathrm{d}\mu^j_{e_a,e_a}(\lambda), \end{aligned}$$

and where

$$\mu^j_{e_a,e_a}(\lambda) \equiv \langle e_a | P^j(\lambda) | e_a \rangle$$

is a *spectral measure*. For instance, if H^j has j eigenvalues $\{\lambda_a,\ a = 1, \dots, j\}$, possibly with repetition, then

$$\mu^j_{e_a,e_a}(\lambda) = \begin{cases} 0 & \text{if } \lambda < \lambda_1, \\ \langle e_a | P^j(\lambda_k) | e_a \rangle & \text{if } \lambda_k \le \lambda < \lambda_{k+1}, k = 1, \dots, j-1, \\ \langle e_a | P^j(\lambda_j) | e_a \rangle & \text{if } \lambda \ge \lambda_j. \end{cases}$$

Knowing the numbers $E^j(g^j)$, $N^j(g^j)$ and the spectral measures $\mu^j_{e_a,e_a}(\lambda)$ for every a, the calculation of the energy-minimizing discrete density matrix γ^j at fixed (ϕ, u) reduces to the scalar problem

$$\inf_{g^j \in \mathfrak{B}} \left\{ E^j(g^j),\, 0 \le g^j \le 1,\, N^j(g^j) = N \right\}, \tag{48}$$

where $\mathfrak{B}$ denotes the space of bounded real-valued Borel functions over the real line.

Numerically, spectral approximation consists of finding a minimizer in equation (48) by applying the Rayleigh-Ritz method over a finite-dimensional subspace $\mathfrak{B}_k$ of $\mathfrak{B}$ spanned by a chosen spectral basis $\{s^k_1, \dots, s^k_k\}$, $k \in \mathbb{N}$. Any basis that spans the space of real-valued bounded measurable functions can be chosen for spectral discretization. In practice, one would choose a basis in which its spectral integral for each $e_a, a \in \mathbb{N}$,

$$\int_{\sigma(H^j)} s^k_q(\lambda) d\mu^j_{e_a,e_a}(\lambda),$$

can be evaluated at a cost that scales better than cubic with respect to the number of electrons in the system.

Let us introduce the subsets

$$\mathcal{K}^{H^j(\phi,u)}_{N,k} = \left\{ \gamma \in \mathcal{K}^{H^j(\phi,u)}_{N} : \gamma = \sum_{q=1}^{k} c^k_q s^k_q(H^j) \right\}. \tag{49}$$

Then the band energy for a density matrix $\gamma \in \mathcal{K}_{N,k}^{H^j(\phi,u)}$ is

$$\begin{aligned} E^j(\gamma) &= E^j\Big(\sum_{q=1}^{k} c_q^k s_q^k\Big) = \mathrm{Tr}(H^j\gamma) \\ &= \sum_{i=1}^{\infty} \int_{\sigma(H^j)} \lambda \sum_{q=1}^{k} c_q^k s_q^k(\lambda)\, \mathrm{d}\mu_{e_i,e_i}^j(\lambda) \\ &= \sum_{q=1}^{k} c_q^k \Bigg\{ \sum_{i=1}^{\infty} \int_{\sigma(H^j)} \lambda s_q^k(\lambda)\, \mathrm{d}\mu_{e_i,e_i}^j(\lambda) \Bigg\} \equiv \sum_{q=1}^{k} c_q^k w_q^{k,j}, \end{aligned} \tag{50}$$

and the number of electrons in the system for $\gamma \in \mathcal{K}_{N,k}^{H^j(\phi,u)}$ is

$$\begin{aligned} N^j(\gamma) &= N^j\Big(\sum_{q=1}^{k} c_q^k s_q^k\Big) = \mathrm{Tr}(\gamma) \\ &= \sum_{i=1}^{\infty} \int_{\sigma(H^j)} \sum_{q=1}^{k} c_q^k s_q^k(\lambda)\, \mathrm{d}\mu_{e_i,e_i}^j(\lambda) \\ &= \sum_{q=1}^{k} c_q^k \Bigg\{ \sum_{i=1}^{\infty} \int_{\sigma(H^j)} s_q^k(\lambda)\, \mathrm{d}\mu_{e_i,e_i}^j(\lambda) \Bigg\} \equiv \sum_{q=1}^{k} c_q^k n_q^{k,j}. \end{aligned} \tag{51}$$

The minimization of the energy function in equation (48) over $\mathfrak{B}_k$ becomes

$$\inf_{\{c_q^k\}\subset\mathbb{R}^k} E^j\Big(\sum_{q=1}^{k} c_q^k s_q^k\Big),$$

subject to the constraints

$$0 \le c_q^k \le 1, \quad \sum_{q=1}^{k} c_q^k n_q^{k,j} = N.$$

Next we give an example of spectral discretization, spectral binning.

5.3.1 Spectral binning

Spectral binning refers to a basis consisting of a collection of disjoint piece-wise constant functions, also known as *simple functions*. The spectral binning basis is defined over a partition of the fixed interval $[\lambda_{\mathrm{LB}}, \lambda_{\mathrm{U}B}]$ into k sub-intervals, or *bins*, $\{t_q^k, q = 0, \ldots, k\}$. We require that $t_0^k = \lambda_{\mathrm{L}B} \le \lambda_{\min}$ and $\lambda_N \le \lambda_{\mathrm{UB}} = t_k^k < \lambda_{\max}$, where $\lambda_{\min}$ and $\lambda_{\max}$ are the minimum and maximum eigenvalues of H^j, respectively. The choice of $(\lambda_{\mathrm{LB}}, \lambda_{\mathrm{U}B})$ must ensure that the space $\mathcal{K}_{N,k}^{H^j(\phi,u)}$ includes the minimizer $\gamma_{\min}$ to the band energy functional $E^j(g^j)$. Let $s_{t_q^k}(\lambda)$ denote the disjoint piece-wise constant characteristic functions defined on the spectrum of $H^j(\phi, u)$,

$$s_{t_q^k}(\lambda) \equiv \begin{cases} 1, & \text{if } t_{q-1}^k \le \lambda \le t_q^k, \\ 0, & \text{otherwise.} \end{cases}$$

We define $\mathfrak{B}_k$ as the collection of constant simple functions $\{s_{t_q^k}\}_{q=1}^k$ associated with this partition. These functions form a natural basis because they are dense over the space of integrable real functions over $[\lambda_{\mathrm{L}B}, \lambda_{\mathrm{U}B}]$. The density matrix $\gamma^j_k \in \mathcal{K}^{H^j(\phi,u)}_{N,k}$ using the spectral theorem in the spectral binning basis is

$$\gamma^j_k = \int_{\sigma(H^j)} \sum_{q=1}^{k} c^k_q s_{t^k_q}(\lambda) \,\mathrm{d}P^j(\lambda). \tag{52}$$

For any $\gamma \in \mathcal{K}^{H^j(\phi,u)}_{N,k}$ with associated coefficients $\{c^k_q\}^k_{q=1}$ as in equation (52), the corresponding band energy is

$$\begin{aligned} E^j(\gamma) = E^j\Big(\sum_{q=1}^{k} c^k_q s_{t^k_q}\Big) &= \mathrm{Tr}(H^j\gamma) \\ &= \sum_{q=1}^{k} c^k_q \left(\sum_{i=1}^{\infty} \int_{\sigma(H^j)} \lambda s_{t^k_q}(\lambda) \,\mathrm{d}\mu^j_{e_i,e_i}(\lambda)\right) = \sum_{q=1}^{k} c^k_q w^{k,j}_q, \end{aligned}$$

and

$$\begin{aligned} N^j(\gamma) = N^j\Big(\sum_{q=1}^{k} c^k_q s_{t^k_q}\Big) &= \mathrm{Tr}(\gamma) \\ &= \sum_{q=1}^{k} c^k_q \left(\sum_{i=1}^{\infty} \int_{\sigma(H^j)} s_{t^k_q}(\lambda) \,\mathrm{d}\mu^j_{e_i,e_i}(\lambda)\right) = \sum_{q=1}^{k} c^k_q n^{k,j}_q, \end{aligned}$$

where $n^{k,j}_q$ can be interpreted as the number of eigenvalues in the interval (t^k_{q-1}, t^k_q), hence giving rise to the name of the method, *spectral binning*.

The minimization over $\mathfrak{B}_k$ in equation (48) becomes a linear programming problem,

$$\inf_{\{c^k_q\}\subset\mathbb{R}^k} \sum_{q=1}^{k} c^k_q w^{k,j}_q, \tag{53}$$

subject to the linear constraints

$$0 \le c^k_q \le 1, \quad \sum_{q=1}^{k} c^k_q n^{k,j}_q = N. \tag{54}$$

To proceed with the spectral binning discretization numerically, we have to evaluate the quantities $\{n^{k,j}_q\}$ and $\{w^{k,j}_q\}$. In the next subsection we explain in more detail how this is done.

5.3.2 Numerical evaluation of $\{n^{k,j}_q\}^k_{q=1}$

By Sylvester's law of inertia [31], $n^{k,j}_q$ equals the number of eigenvalues of $H^j(\phi, u)$ contained in the sub-interval (t^k_{q-1}, t^k_q). The inertia of a given matrix H^j is denoted by the number triple $(\mathcal{N}_-, \mathcal{N}_0, \mathcal{N}_+)$, where $\mathcal{N}_-$ denotes the number of negative eigenvalues of H, $\mathcal{N}_0$ the dimension of the kernel of H, and $\mathcal{N}_+$ the number of positive eigenvalues of H^j. Sylvester proved that the inertia of a matrix is invariant under congruent transformations of the matrix.

The congruent transformation we adopt is the decomposition $H^j = LDL^T$, where D is a diagonal matrix and L is a lower triangular matrix. The number of negative elements in D corresponds to the number of negative eigenvalues of the matrix H^j, [21]. To find the number of eigenvalues of the discrete Hamiltonian matrix H^j in an interval $[t^k_{q-1}, t^k_q]$, we need to perform the LDL^T decomposition twice,

$$\begin{aligned} H^j - t^k_{q-1}\mathcal{I}^j &= L_{t^k_{q-1}} D_{t^k_{q-1}} L^T_{t^k_{q-1}}, \\ H^j - t^k_q\mathcal{I}^j &= L_{t^k_q} D_{t^k_q} L^T_{t^k_q}. \end{aligned} \tag{55}$$

Here, $\mathcal{I}^j$ denotes the $j \times j$ identity matrix. For a non-orthogonal spatial discretization, we simply replace $\mathcal{I}^j$ with the corresponding mass matrix $\mathcal{M}^j$. Let $\mathcal{N}_-(D_{t^k_q})$ denote the number of negative eigenvalues of $D_{t^k_q}$. Then it holds

$$n^k_q = \mathcal{N}_-(D_{t^k_q}) - \mathcal{N}_-(D_{t^k_{q-1}}).$$

Considering the computational cost for the LDL^T decomposition, for a $j \times j$ matrix with half bandwidth W, the number of operations for the LDL^T decomposition is, see e.g. [21],

$$C_{LDL^T} = \frac{W(W+1)j}{2}. \tag{56}$$

Based on equation (56), for k partitions or 'bins' of the spectrum, the total number of operations to obtain the number of eigenvalues in each bin is

$$C_{\text{binning}} = \frac{W(W+1)kj}{2}. \tag{57}$$

However, the half bandwidth W of the Hamiltonian scales with respect to the number of spatial discretizations depending on the spatial dimension of the system. According to [20], the computational cost for the LDL^T decomposition of a molecular system in 3D at worst scales like N^2. Note that by (57), the computational cost of the binning method scales linearly with respect to the number of spectral discretizations k.

5.4 Numerical evaluation of $\{w^{k,j}_q\}^k_{q=1}$

Unlike $n^{k,j}_q$ introduced in (51), we cannot evaluate $w^{k,j}_q$ defined in (50) directly at a cost that scales better than cubic with respect to the number of electrons in the system. Hence we proceed to make one more approximation. Let $\{m^k_q\}^k_{q=1}$ be the center of mass of each partition, defined by

$$m^k_q \equiv \frac{w^{k,j}_q}{n^{k,j}_q} = \frac{1}{n^{k,j}_q} \sum_{i=1}^{\infty} \left(\int_{\sigma(H^j)} \lambda s_{t^k_q}(\lambda)\, \mathrm{d}\mu^j_{e_i,e_i}(\lambda) \right). \tag{58}$$

We approximate the center of mass m^k_q in the interval (t^k_{q-1}, t^k_q) by

$$m^k_q \approx \frac{t^k_q - t^k_{q-1}}{2}. \tag{59}$$

This approximation implies the spectral approximation of the band energy as

$$\begin{aligned} \mathrm{Tr}\big(H^j(\phi,u)\gamma^j\big) &= \sum_{i=1}^{\infty} \int_{\sigma(H^j)} \sum_{q=1}^{k} c_q \lambda s_{t^k_q}(\lambda)\, \mathrm{d}\mu^j_{e_i,e_i}(\lambda) \\ &\approx \sum_{q=1}^{k} c_q m^{k,j}_q n^{k,j}_q \equiv \tilde{\mathrm{Tr}}(H^j(\phi,u)\gamma^j). \end{aligned} \tag{60}$$

This approximation of $\{w^{k,j}_q\}^k_{q=1}$ introduces an error over the Rayleigh-Ritz approximation of the discrete band energy. However, in the next section we are going to show that this error is controllable.

6 Convergence with respect to spectral and spatial discretization

We define the relevant functionals so that we can best utilize the machineries of Γ-convergence.

Part I: Definition of the limit functionals.

Starting from equation (39), we consider the minimization problem

$$\varepsilon_{\mathrm{GS}}^{\mathrm{R}EKS} = \inf_{u \in \mathcal{U}} T(u),$$

where $T : \mathcal{U} \to \mathbb{R}$ is defined by

$$T(u) = B_{\mathrm{xc}}^{*}(u) + \sup_{\phi \in \mathcal{V}} S(u, \phi)$$

and $S(u, \cdot) : \mathcal{V} \to \mathbb{R}$ is

$$S(u, \phi) = -\int_{\Omega} \big(C_S |\nabla \phi(\mathbf{r})|^2 - b(\mathbf{r}, \{\mathbf{R}\})\phi(\mathbf{r}) \big)\, \mathrm{d}\mathbf{r} + \inf_{\gamma \in \mathcal{X}} \Big\{ E_{\mathrm{band}}(u, \phi, \gamma) + I_{\mathcal{K}_N^{H(\phi,u)}}(\gamma) \Big\}. \tag{61}$$

Here, $I_{\mathcal{M}}$ for a set $\mathcal{M}$ denotes the indicator function of convex analysis,

$$I_{\mathcal{M}}(u) \equiv \begin{cases} 0 & \text{if } u \in \mathcal{M}, \\ +\infty & \text{otherwise.} \end{cases}$$

In (61), the minimization over $\mathcal{K}_N$ was replaced by the minimization over $\mathcal{K}_N^{H(\phi,u)}$. This ensures the existence of a spectral function and was justified in equation (42).

Part II: Definition of the functionals with combined spectral and spatial approximation.

For $j \in \mathbb{N}$, based on the identity (39), we introduce the family of energies

$$\varepsilon_{j,k_j} = \inf_{u \in \mathcal{U}} T_{j,k_j}(u),$$

where $T_{j,k_j} : \mathcal{U} \to \mathbb{R} \cup \{+\infty\}$ are defined by

$$T_{j,k_j}(u) = B_{\mathrm{xc}}^{*}(u) + \sup_{\phi \in \mathcal{V}} S^{j,k_j}(u, \phi) + I_{\mathcal{U}_j}(u)$$

and $S^{j,k_j}(u, \cdot) : \mathcal{V} \to \mathbb{R} \cup \{-\infty\}$ are given by

$$S^{j,k_j}(u, \phi) = -\int_{\Omega} \big(C_S |\nabla \phi(\mathbf{r})|^2 - b(\mathbf{r}, \{\mathbf{R}\})\phi(r) \big)\, \mathrm{d}\mathbf{r} + \inf_{\gamma \in \mathcal{X}} \Big\{ E_{\mathrm{band}_{j,k_j}}(u, \phi, \gamma) + I_{\mathcal{K}_{N,k_j}^{H^j(\phi,u)}}(\gamma) \Big\} - I_{\mathcal{V}_j}(\phi). \tag{62}$$

In (62), we introduced the approximated constrained sets of density matrices

$$\mathcal{K}_{N,k_j}^{H^j(\phi,u)} = \Big\{ \gamma \in \mathcal{K}_N : \gamma = \sum_{i=1}^{k_j} c_i^{k_j} s_{t_i^{k_j}}(H^j),\ 0 \le c_i^{k_j} \le 1 \Big\}$$

and the discrete band energies $E_{\mathrm{band}_{j,k_j}}(u, \phi, \cdot) : \mathcal{X} \to \mathbb{R}$,

$$E_{\mathrm{band}_{j,k_j}}(u, \phi, \gamma) = \tilde{\mathrm{Tr}}\big(H^j(\phi, u)\gamma \big), \tag{63}$$

where $\tilde{\mathrm{Tr}}(\cdot)$ (depending on k_j) is the approximation of the trace operator described in equation (60). We emphasize that this is the actual numerical approximation of the binning algorithm introduced in Section 5.4.

Summarizing (50) and (60), for $\gamma_{k_j} \in \mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}$ the approximate trace operator is

$$\begin{aligned}\tilde{\operatorname{Tr}}(H^j\gamma_{k_j}) &= \sum_{i=1}^{\infty}\sum_{q=1}^{k_j} c_q^{k_j} m_q^{k_j} \int_{t_q^{k_j}}^{t_{q+1}^{k_j}} s_q^{k_j}(\lambda)\,\mathrm{d}\mu_{e_i,e_i}(\lambda) \\ &= \sum_{i=1}^{\infty}\sum_{q=1}^{k_j} c_q^{k_j} m_q^{k_j} \big(\mu_{e_i,e_i}(t_{q+1}^{k_j}) - \mu_{e_i,e_i}(t_q^{k_j})\big), \qquad (64)\end{aligned}$$

where $m_q^{k_j} \equiv \frac{t_{q+1}^{k_j}+t_q^{k_j}}{2}$ denotes as in (59) the arithmetic mean.

We show convergence w.r.t. both spectral and spatial discretization using three nested Γ-convergence proofs. We first establish the convergence of the exact band energies $\operatorname{Tr}(H^j(\phi_j,u_j)\gamma_j)$. Then, in Section 6.2, we validate the convergence of the approximate trace operators.

6.1 The Γ-convergence of the exact band energies $\operatorname{Tr}\big(H^j(\phi_j,u_j)\gamma_j\big)$

Lemma 6.1 *If $u_j \rightharpoonup u$ in $\mathcal{U}$ and $\phi_j \rightharpoonup \phi$ in $\mathcal{V}$, then*

$$\liminf_{j\to\infty}\Big\{\operatorname{Tr}\big(H^j(u_j,\phi_j)\gamma_j\big) + I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma)\Big\} \geq E_{\text{band}}(u,\phi,\gamma) + I_{\mathcal{K}_N^{H(\phi,u)}}(\gamma)$$

for every $\gamma \in \mathcal{X}$ and for all $\gamma_j \overset{*}{\rightharpoonup} \gamma$ in $\mathcal{X}$.

Proof We consider four disjoint cases.

1. Let $\gamma \in \mathcal{K}_N^{H^j(\phi,u)}$ and $\{\gamma_j\}_{j\in\mathbb{N}} \subset \mathcal{X}$ be a sequence with $\gamma_j \overset{*}{\rightharpoonup} \gamma$ such that there exists a $q_1 \in \mathbb{N}$ so that $\gamma_j \in \mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}$ for all $j \geq q_1$.

   By the lower semi-continuity of the kinetic energy proved in Lemma 4.6,

   $$\liminf_{j\to\infty} \operatorname{Tr}(-\Delta\gamma_j) \geq \operatorname{Tr}(-\Delta\gamma),$$

   and by the compact embedding of $\mathcal{W}_0^{1,2}(\Omega)$ in $\mathcal{L}^2(\Omega)$, $\gamma_j \overset{*}{\rightharpoonup} \gamma$ implies that $\rho_{\gamma_j} \to \rho_\gamma$ in $\mathcal{L}^2(\Omega)$. This yields

   $$\begin{aligned}\lim_{j\to\infty} \operatorname{Tr}\big((\Phi_j - U_j)\gamma_j\big) &= \lim_{j\to\infty}\int_\Omega \big(\phi_j(\mathbf{r}) - u_j(\mathbf{r})\big)\rho_{\gamma_j}(\mathbf{r})\,\mathrm{d}\mathbf{r} = \int_\Omega \big(\phi(\mathbf{r}) - u(\mathbf{r})\big)\rho_\gamma(\mathbf{r})\,\mathrm{d}\mathbf{r} \\ &= \operatorname{Tr}\big((\Phi - U)\gamma\big),\end{aligned}$$

   leading to

   $$\liminf_{j\to\infty} \operatorname{Tr}\big(H^j(\phi_j,u_j)\gamma_j\big) \geq \operatorname{Tr}\big(H(\phi,u)\gamma\big).$$

2. Let $\gamma \in \mathcal{K}_N^{H(\phi,u)}$ and $\{\gamma_j\}_{j\in\mathbb{N}} \subset \mathcal{X}$ be a sequence such that there exists a $q_2 \in \mathbb{N}$ so that $\gamma_j \notin \mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}$ for all $j \geq q_2$.

   In this case we have trivially

   $$\liminf_{j\to\infty}\Big\{\operatorname{Tr}\big(H^j(u_j,\phi_j)\gamma_j\big) + I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma)\Big\} = +\infty \geq E_{\text{band}}(u,\phi,\gamma) + I_{\mathcal{K}_N^{H(\phi,u)}}(\gamma).$$

3. Let $\gamma \notin \mathcal{K}_N^{H(\phi,u)}$ and $\{\gamma_j\}_{j\in\mathbb{N}} \subset \mathcal{X}$ be a sequence such that there exists a $q_3 \in \mathbb{N}$ so that $\gamma_j \notin \mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}$ for all $j \geq q_3$.

   In this case we have trivially

$$\liminf_{j\to\infty} \Big\{ \mathrm{Tr}\big(H^j(u_j,\phi_j)\gamma_j\big) + I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma) \Big\} = E_{\mathrm{band}}(u,\phi,\gamma) + I_{\mathcal{K}_N^{H(\phi,u)}}(\gamma) = +\infty.$$

4. Now we show that if $\gamma \notin \mathcal{K}_N^{H(\phi,u)}$, then there cannot exist a sequence $\gamma_j \overset{*}{\rightharpoonup} \gamma$ such that there exists a $q_4 \in \mathbb{N}$ so that $\gamma_j \in \mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}$ for all $j \geq q_4$.

   Let $\{\xi_i\}_{i\in\mathbb{N}} \subset \mathcal{W}_0^{1,2}(\Omega)$ represent the eigenvectors of $H(\phi,u)$ which are known to form an orthonormal basis of $\mathcal{L}^2(\Omega)$. Similarly, for $j \in \mathbb{N}$, let $\{\xi_i^j\}_{i\in\mathbb{N}} \subset \mathcal{W}_0^{1,2}(\Omega)$ be the eigenvectors of $H^j(\phi_j,u_j)$. From the Rayleigh-Ritz discretization of the Hamiltonian, we can ensure the convergence of the eigenvectors, i.e. for every $i \in \mathbb{N}$,

$$\lim_{j\to\infty} \|\xi_i^j - \xi_i\|_{\mathcal{L}^2(\Omega)} = 0, \qquad \lim_{j\to\infty} \xi_i^j = \xi_i.$$

   Since $\gamma \notin \mathcal{K}_N^{H(\phi,u)}$, for the case considered here, there must exist an eigenvector of H which is not an eigenvector of γ. Let us denote it by ξ_1. So it holds

$$\gamma\xi_1 = \sum_{q=1}^{\infty} c_{1q}\xi_q,$$

   and there must exist an index $p \in \mathbb{N}$, $p \neq 1$, such that $c_{1p} \neq 0$. Consider this c_{1p}. Then

$$c_{1p} = \langle \gamma\xi_1, \xi_p \rangle = \lim_{j\to\infty} \langle \gamma_j\xi_1, \xi_p \rangle = \lim_{j\to\infty} \langle g_j(H^j)\xi_1, \xi_p \rangle.$$

   Therefore, for $p \neq 1$,

$$\begin{aligned} \lim_{j\to\infty} \langle g_j(H^j)\xi_1, \xi_p \rangle &= \lim_{j\to\infty} \langle g_j(H^j)\xi_1 - \xi_1^j + \xi_1^j, \xi_p \rangle \\ &= \lim_{j\to\infty} \langle g_j(H^j)\xi_1^j, \xi_p \rangle + \lim_{j\to\infty} \langle g_j(H^j)(\xi_1 - \xi_1^j), \xi_p \rangle \\ &= \lim_{j\to\infty} g_j(\lambda_1^j)\langle \xi_1^j, \xi_p \rangle = 0. \end{aligned}$$

   We then have $c_{1p} = 0$ for all $p \neq 1$, contradicting our assumption. Hence we have shown that if $\gamma \notin \mathcal{K}_N^{H(\phi,u)}$, there cannot be a sequence $\{\gamma_j\}_{j\in\mathbb{N}}$ with $\gamma_j \in \mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}$ for all $j \in \mathbb{N}$ and $\gamma_j \overset{*}{\rightharpoonup} \gamma$.

The above four cases demonstrate that for all $\gamma \in \mathcal{X}$ and for all $\gamma_j \overset{*}{\rightharpoonup} \gamma$ in $\mathcal{X}$,

$$\liminf_{j\to\infty} \Big\{ \mathrm{Tr}\big(H^j(u_j,\phi_j)\gamma_j\big) + I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma) \Big\} \geq E_{\mathrm{band}}(u,\phi,\gamma) + I_{\mathcal{K}_N^{H(\phi,u)}}(\gamma). \qquad \square$$

Lemma 6.2 *Let $u_j \rightharpoonup u$ in $\mathcal{U}$ and $\phi_j \rightharpoonup \phi$ in $\mathcal{V}$. Then for all $\gamma \in \mathcal{K}_N^{H(\phi,u)}$, there exists a recovery sequence $\gamma_j \overset{*}{\rightharpoonup} \gamma$ such that*

$$\limsup_{j\to\infty} \operatorname{Tr}\big(H^j(u_j,\phi_j)\gamma_j\big) \le E_{\text{band}}(u,\phi,\gamma)$$

and

$$\operatorname{Tr}\big(H^j(u_j,\phi_j)\gamma\big) + I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma) \xrightarrow{\Gamma} E_{\text{band}}(u,\phi,\gamma) + I_{\mathcal{K}_N^{H(\phi,u)}}(\gamma)$$

with respect to the weak-topology in $\mathcal{X}$ as $j \to \infty$.*

Proof We consider two disjoint cases.

1. If $\gamma \notin \mathcal{K}_N^{H(\phi,u)}$, then let the recovery sequence be defined by the finite-rank operators that converge to γ in $\|\cdot\|_{\mathcal{X}}$. This sequence of finite-rank operators exists due to the Rayleigh-Ritz method and is dense in $\mathcal{X}$. With this recovery sequence, it trivially holds

$$\limsup_{j\to\infty} \operatorname{Tr}\big(H^j(u_j,\phi_j)\gamma_j\big) \le E_{\text{band}}(u,\phi,\gamma) = +\infty.$$

2. If $\gamma \in \mathcal{K}_N^{H(\phi,u)}$, then without loss of generality, we write

$$\gamma = \sum_{i=1}^{\infty} 2\alpha_i \xi_i\rangle\langle\xi_i, \tag{65}$$

where $\{\xi_i\}_{i\in\mathbb{N}}$, $\{\xi_i^j\}_{i\in\mathbb{N}}$ denote the sets of eigenvectors of $H(\phi,u)$ and $H^j(\phi_j,u_j)$, respectively, as in Lemma 6.1.

Let us define the sequence of finite-rank operators,

$$\gamma_j = \sum_{i=1}^{j} 2\alpha_i \xi_i^j\rangle\langle\xi_i^j. \tag{66}$$

We proceed to show that $\gamma_j \to \gamma$ w.r.t. $\|\cdot\|_{\mathcal{X}}$. From Theorem VI.10 in [24], there exists an unique partial isometry P such that

$$|\gamma - \gamma_j| = P(\gamma - \gamma_j). \tag{67}$$

Now we show the strong convergence of $\gamma_j \to \gamma$ in the norm sense of $\mathcal{X}$ as follows. Utilizing equation (67), the dual operator P^* of P, the Cauchy-Schwarz inequality and the fact that both P and P^* are isometries, we find

$$\begin{aligned}\lim_{j\to\infty}\mathrm{Tr}(|\gamma-\gamma_j|) &= \lim_{j\to\infty}\mathrm{Tr}(P(\gamma-\gamma_j))\\ &= \lim_{j\to\infty}\sum_{p=1}^{\infty}\langle P(\gamma-\gamma_j)\xi_p,\xi_p\rangle\\ &= \lim_{j\to\infty}\sum_{p=1}^{\infty}\langle(\gamma-\gamma_j)\xi_p,P^*\xi_p\rangle\\ &\le \lim_{j\to\infty}\sum_{p=1}^{\infty}\|(\gamma-\gamma_j)\xi_p\|_{\mathcal{L}^2(\Omega)}\|P^*\xi_p\|_{\mathcal{L}^2(\Omega)}\\ &\le \lim_{j\to\infty}\sum_{p=1}^{\infty}\|(\gamma-\gamma_j)\xi_p\|_{\mathcal{L}^2(\Omega)}\|\xi_p\|_{\mathcal{L}^2(\Omega)}\\ &= \lim_{j\to\infty}\sum_{p=1}^{\infty}\|(\gamma-\gamma_j)\xi_p\|_{\mathcal{L}^2(\Omega)}. \end{aligned}\tag{68}$$

Let us consider just one of the terms in equation (68) for fixed summation index p. We now look at its projection onto the eigen-basis $\{\xi_i\}_{i\in\mathbb{N}}$ and find with (65), (66)

$$\begin{aligned}\lim_{j\to\infty}\|(\gamma-\gamma_j)\xi_p\|^2_{\mathcal{L}^2(\Omega)} &= \lim_{j\to\infty}\sum_{q=1}^{\infty}\Big|\langle(\gamma-\gamma_j)\xi_p,\xi_q\rangle\Big|^2\\ &= \lim_{j\to\infty}\left\{\sum_{q=1}^{\infty}\left|2\alpha_q\langle\xi_p,\xi_q\rangle-\sum_{i=1}^{j}2\alpha_i\langle\xi_p,\xi_i^j\rangle\langle\xi_i^j,\xi_q\rangle\right|^2\right\}\\ &\le \lim_{j\to\infty}\left\{\left|2\alpha_p-\sum_{i=1}^{j}2\alpha_i\langle\xi_p,\xi_i^j\rangle^2\right|^2+\sum_{q=1,q\neq p}^{\infty}\left|\sum_{i=1}^{j}2\alpha_i\langle\xi_p,\xi_i^j\rangle\langle\xi_i^j,\xi_q\rangle\right|^2\right\}\\ &\le \lim_{j\to\infty}\left\{\left|2\alpha_p-\sum_{i=1}^{j}2\alpha_i\langle\xi_p,(\xi_i^j-\xi_i)+\xi_i\rangle^2\right|^2\right.\\ &\qquad\left.+\sum_{q=1,q\neq p}^{\infty}\left|\sum_{i=1}^{j}2\alpha_i\langle\xi_p,(\xi_i^j-\xi_i)+\xi_i\rangle\langle(\xi_i^j-\xi_i)+\xi_i,\xi_q\rangle\right|^2\right\}=0. \end{aligned}\tag{69}$$

The above limit converges to 0 since for every $q\in\mathbb{N}$

$$\lim_{j\to\infty}\langle\xi_i,\xi_q^j-\xi_q\rangle=\lim_{j\to\infty}\langle\xi_q^j-\xi_q,\xi_i\rangle=0.$$

With the help of (69), we find

$$0=\sum_{p=1}^{\infty}\liminf_{j\to\infty}\|(\gamma-\gamma_j)\xi_p\|_{\mathcal{L}^2(\Omega)}\le\liminf_{j\to\infty}\sum_{p=1}^{\infty}\|(\gamma-\gamma_j)\xi_p\|_{\mathcal{L}^2(\Omega)}.\tag{70}$$

Similarly, by Jensen's inequality,

$$\limsup_{j\to\infty}\sum_{p=1}^{\infty}\|(\gamma-\gamma_j)\xi_p\|_{\mathcal{L}^2(\Omega)}\le\sum_{p=1}^{\infty}\limsup_{j\to\infty}\|(\gamma-\gamma_j)\xi_p\|_{\mathcal{L}^2(\Omega)}=0.\tag{71}$$

As a result of (70), (71) we have $0 \leq \liminf_{j\to\infty} \|(\gamma-\gamma_j)\xi_p\|_{\mathcal{L}^2(\Omega)} \leq \limsup_{j\to\infty} \|(\gamma-\gamma_j)\xi_p\|_{\mathcal{L}^2(\Omega)} \leq 0$, implying

$$\lim_{j\to\infty} \mathrm{Tr}(|\gamma-\gamma_j|) \leq \lim_{j\to\infty} \ \|(\gamma-\gamma_j)\xi_p\|_{\mathcal{L}^2(\Omega)} = 0.$$

Now we are going to use spectral theory to approximate each γ_j. By the choice of γ_j, there are suitable bounded Borel functions g_j such that

$$\gamma_j = g_j(H^j).$$

Next we define the sequence $\tilde{\gamma}_{j,k}$ by

$$\tilde{\gamma}_{j,k} = \sum_{i=1}^{k} c_i^{k,j}\, s_{t_i^k}(H^j),$$

where

$$c_i^{k,j} \equiv \max\{g_j(t_i^k), g_j(t_{i+1}^k)\},$$

and $\{t_1^k, \dots, t_k^k\}$ is the partition of the interval $[\lambda_{\mathrm{L}B}, \lambda_{\mathrm{U}B}]$ introduced in Section 5.3.1. We can show that for every $j \in \mathbb{N}$

$$\mathrm{Tr}(|\tilde{\gamma}_{j,k} - \gamma_j|) \to 0 \tag{72}$$

as $k \to \infty$, see Theorem 2.29 in [34]. However, the trace of $\tilde{\gamma}_{j,k}$ does not satisfy the trace condition for every k, i.e.

$$\mathrm{Tr}(\tilde{\gamma}_{j,k}) \neq N.$$

Nevertheless, since

$$\lim_{k\to\infty} \mathrm{Tr}(\tilde{\gamma}_{j,k}) = N,$$

we can normalize the trace to N by introducing

$$\gamma_{j,k} \equiv \frac{N}{\mathrm{Tr}(\tilde{\gamma}_{j,k})}\tilde{\gamma}_{j,k},$$

Here, due to (72), we may assume $\mathrm{Tr}(\tilde{\gamma}_{j,k}) \neq 0$ for all j and k.

In conclusion, we have

$$\lim_{k\to\infty} \mathrm{Tr}(|\gamma_{j,k} - \gamma_j|) \leq \lim_{k\to\infty} \left\{\mathrm{Tr}(|\gamma_{j,k} - \tilde{\gamma}_{j,k}|) + \mathrm{Tr}(|\tilde{\gamma}_{j,k} - \gamma_j|)\right\} = 0. \tag{73}$$

Eqn. (73) implies that for every j there is an index $k_j \in \mathbb{N}$, $k_j \to \infty$ as $j \to \infty$, such that

$$\mathrm{Tr}(|\gamma_{k_j} - \gamma_j|) \leq \frac{1}{j}.$$

Hence the recovery sequence for every $\gamma \in \mathcal{K}_N^{H(\phi,u)}$ can be defined as $\gamma_{k_j} \in \mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}$, and

$$\begin{aligned}\lim_{j\to\infty} \mathrm{Tr}(|\gamma_{k_j} - \gamma|) &\leq \lim_{j\to\infty} \{\mathrm{Tr}(|\gamma_{k_j} - \gamma_j|) + \mathrm{Tr}(|\gamma_j - \gamma|)\} \\ &\leq \lim_{j\to\infty} \left\{\frac{1}{j} + \mathrm{Tr}(|\gamma_j - \gamma|)\right\} = 0.\end{aligned}$$

Now, in order to show that

$$\mathrm{Tr}\big(\big||\nabla|(\gamma_{k_j}-\gamma)|\nabla|\big|\big)\to 0$$

as $j\to\infty$, we use that $(\gamma_{k_j}-\gamma)\in\mathcal{X}$ and

$$\lim_{j\to\infty}\|\gamma_{k_j}-\gamma\|_{\sup}\le\lim_{j\to\infty}\mathrm{Tr}(|\gamma_{k_j}-\gamma|)=0.$$

Combining the above arguments, it follows

$$\begin{aligned}\liminf_{j\to\infty}\mathrm{Tr}\big(\big||\nabla|(\gamma_{k_j}-\gamma)|\nabla|\big|\big)&=\liminf_{j\to\infty}\mathrm{Tr}(P|\nabla|(\gamma_{k_j}-\gamma)|\nabla|)\\&=\liminf_{j\to\infty}\sum_{q=1}^{\infty}\langle P|\nabla|(\gamma_{k_j}-\gamma)|\nabla|\xi_q,\xi_q\rangle\\&\ge\sum_{q=1}^{\infty}\liminf_{j\to\infty}\langle(\gamma_{k_j}-\gamma)|\nabla|\xi_q,|\nabla|P^*\xi_q\rangle=0,\end{aligned}\tag{74}$$

and similarly

$$\begin{aligned}\limsup_{j\to\infty}\mathrm{Tr}\big(\big||\nabla|(\gamma_{k_j}-\gamma)|\nabla|\big|\big)&=\limsup_{j\to\infty}\mathrm{Tr}(P|\nabla|(\gamma_{k_j}-\gamma)|\nabla|)\\&=\limsup_{j\to\infty}\sum_{q=1}^{\infty}\langle P|\nabla|(\gamma_{k_j}-\gamma)|\nabla|\xi_q,\xi_q\rangle\\&\le\sum_{q=1}^{\infty}\limsup_{j\to\infty}\langle(\gamma_{k_j}-\gamma)|\nabla|\xi_q,|\nabla|P^*\xi_q\rangle=0.\end{aligned}\tag{75}$$

Together, (74) and (75) yield

$$\lim_{j\to\infty}\mathrm{Tr}\big(\big||\nabla|(\gamma_{k_j}-\gamma)|\nabla|\big|\big)=0.$$

So we have shown that for indices (j,k_j), we can choose $\gamma_{k_j}\in\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}$ as the recovery sequence and $\gamma_{k_j}\to\gamma\in\mathcal{K}_N^{H(\phi,u)}$. For this sequence, the band energy converges in the limit,

$$\limsup_{j\to\infty}\mathrm{Tr}\big(H^j(u_j,\phi_j)\gamma_j\big)=E_{\mathrm{band}}(u,\phi,\gamma),$$

where $\gamma\in\mathcal{K}_N^{H(\phi,u)}$, $\phi_j\rightharpoonup\phi$ in $\mathcal{V}$ and $u_j\rightharpoonup u$ in $\mathcal{U}$.

Together, the above two cases prove that the limsup condition is satisfied and that in the limit $j\to\infty$

$$\mathrm{Tr}\big(H^j(u_j,\phi_j)\gamma\big)+I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma)\xrightarrow{\Gamma}E_{\mathrm{band}}(u,\phi,\gamma)+I_{\mathcal{K}_N^{H(\phi,u)}}(\gamma).\qquad\square$$

Lemma 6.3 *For every $\phi_j\rightharpoonup\phi$ in $\mathcal{V}$ and every $u_j\rightharpoonup u$ in $\mathcal{U}$, the family of functionals*

$$\Big\{\mathrm{Tr}\big(H^j(u_j,\phi_j)\gamma\big)+I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma)\Big\}_{j\in\mathbb{N}}$$

is equi-coercive with respect to the weak-topology in $\mathcal{X}$.*

Proof This proof is similar to the proof of Lemma 5.1. It is reproduced here for the sake of completeness. For every $\gamma \in \mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}$, we have the bounds from below

$$\begin{aligned}
\mathrm{Tr}\big(H^j(u_j,\phi_j)\gamma\big) &= \frac{1}{2}\mathrm{Tr}(-\Delta\gamma) + \mathrm{Tr}(\Phi_j\gamma) - \mathrm{Tr}(U_j\gamma) \\
&\geq \frac{1}{2}\mathrm{Tr}(-\Delta\gamma) - (\|\phi_j\|_{\mathcal{L}^2(\Omega)} + \|u_j\|_{\mathcal{U}})\|\rho_\gamma\|_{\mathcal{L}^2(\Omega)} \\
&\geq \frac{1}{2}\mathrm{Tr}(-\Delta\gamma) - C_{10}(\|\phi\|_{\mathcal{L}^2(\Omega)} + \|u_j\|_{\mathcal{L}^2(\Omega)})\|\rho_\gamma\|_{\mathcal{L}^1(\Omega)}^{\frac{1}{4}}\|\rho_\gamma\|_{\mathcal{L}^3(\Omega)}^{\frac{3}{4}} \qquad (76)\\
&\geq \frac{1}{2}\mathrm{Tr}(-\Delta\gamma) - C_{11}(\|\phi_j\|_{\mathcal{L}^2(\Omega)} + \|u_j\|_{\mathcal{L}^2(\Omega)})N^{1/4}\|\nabla\sqrt{\rho_\gamma}\|_{\mathcal{L}^2(\Omega)}^{\frac{3}{2}} \qquad (77)\\
&\geq \frac{1}{2}\mathrm{Tr}(-\Delta\gamma) - C_{12}\|\nabla\sqrt{\rho_\gamma}\|_{\mathcal{L}^2(\Omega)}^{\frac{3}{2}}, \qquad (78)
\end{aligned}$$

where interpolation inequalities are used to obtain (76), the Gagliardo–Nirenberg–Sobolev inequality is used to obtain (77), and with the constant

$$C_{12} \equiv C_{11}\sup_{j\in\mathbb{N}}\Big\{\|\phi_j\|_{\mathcal{L}^2(\Omega)} + \|u_j\|_{\mathcal{L}^2(\Omega)}\Big\}N^{1/4}.$$

Since

$$\mathrm{Tr}(-\Delta\gamma) \geq \|\nabla\sqrt{\rho_\gamma}\|_{\mathcal{L}^2(\Omega)}^2,$$

the kinetic energy is the dominating term in the inequality. Hence, for any $t \in \mathbb{R}$ the level sets

$$\Big\{\gamma \in \mathcal{X} \,:\, \mathrm{Tr}\big(H^j(u_j,\phi_j)\gamma\big) + I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma) \leq t\Big\}$$

are bounded,

$$t \geq \frac{1}{2}\|\gamma\|_{\mathcal{X}} - C_{12}\|\sqrt{\rho_\gamma}\|_{\mathcal{L}^2(\Omega)}^{\frac{3}{2}} - \frac{N}{2}.$$

By the results in [15], this shows that for every j and k_j, the level sets of $\Big\{\mathrm{Tr}\big(H^j(u_j,\phi_j)\,\cdot\,\big) + I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma)\Big\}$ are precompact and hence equi-coercive. □

Lemma 6.4 *If $\phi_j \rightharpoonup \phi$ in $\mathcal{V}$ and $u_j \rightharpoonup u$ in $\mathcal{U}$, then*

$$\lim_{j\to\infty}\inf_{\gamma\in\mathcal{X}}\Big\{\mathrm{Tr}\big(H^j(u_j,\phi_j)\gamma\big) + I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma)\Big\} = \inf_{\gamma\in\mathcal{X}}\Big\{E_{\mathrm{band}}(u,\phi,\gamma) + I_{\mathcal{K}_N^{H(\phi,u)}}(\gamma)\Big\}.$$

Proof This is proven using Theorem 7.8 in [12], Lemma 6.2 and Lemma 6.3. □

6.2 Γ-convergence of $E_{\mathrm{band}_{j,k_j}}$ with approximation of the trace operator

In the last section, the Γ-convergence of the exact band energies has been shown. Subsequently, we extend these convergence results to $E_{\mathrm{band}_{j,k_j}}$ introduced in (63), i.e. to the evaluation operators actually used in the binning algorithm.

Lemma 6.5 *Let $u_j \rightharpoonup u$ in $\mathcal{U}$, $\phi_j \rightharpoonup \phi$ in $\mathcal{V}$ as $j \to \infty$ and $\gamma_{k_j} \in \mathcal{K}_{N,k_j}^{H^j}$ for all $j \in \mathbb{N}$. Then*

$$\lim_{j\to\infty}\big|\tilde{\mathrm{Tr}}(H^j\gamma_{k_j}) - \mathrm{Tr}(H^j\gamma_{k_j})\big| = 0. \qquad (79)$$

Proof By direct estimates we find

$$\begin{aligned}
\left|\tilde{\mathrm{Tr}}(H^j\gamma_{k_j}) - \mathrm{Tr}(H^j\gamma_{k_j})\right| &= \left|\sum_{i=1}^{\infty}\sum_{q=1}^{k_j}\int_{t_q^{k_j}}^{t_{q+1}^{k_j}} c_q^{k_j}(m_q^{k_j}-\lambda)s_q^{k_j}(\lambda)\,\mathrm{d}\mu_{e_i,e_i}(\lambda)\right| \\
&= \left|\sum_{i=1}^{\infty}\sum_{q=1}^{k_j} c_q^{k_j}(m_q^{k_j}-\nu_{q,i}^{k_j})\int_{t_q^{k_j}}^{t_{q+1}^{k_j}} s_q^{k_j}(\lambda)\,\mathrm{d}\mu_{e_i,e_i}(\lambda)\right| \qquad (80)\\
&= \left|\sum_{i=1}^{\infty}\sum_{q=1}^{k_j} c_q^{k_j}(m_q^{k_j}-\nu_{q,i}^{k_j})\big(\mu_{e_i,e_i}(t_{q+1}^{k_j}) - \mu_{e_i,e_i}(t_q^{k_j})\big)\right| \\
&\le \left|\sum_{q=1}^{k_j} c_q^{k_j}\frac{h_{k_j}}{2}\sum_{i=1}^{\infty}\big(\mu_{e_i,e_i}(t_{q+1}^{k_j}) - \mu_{e_i,e_i}(t_q^{k_j})\big)\right| \\
&= \left|\sum_{q=1}^{k_j} c_q^{k_j}\frac{h_{k_j}}{2}n_q^{k_j}\right|, \qquad (81)
\end{aligned}$$

where $h_{k_j} := \max_{1\le l\le t_j-1}|t_l^{k_j} - t_{l+1}^{k_j}|$ are the widths of the binning intervals. The numbers $\nu_{q,i}^{k_j} \in (t_q^{k_j}, t_{q+1}^{k_j})$ in equation (80) appear as a result of the mean value theorem for Riemann-Stieltjes integrals with respect to each measure $\mu_{e_i,e_i}(\lambda)$, see e.g. [34].

For each $\epsilon > 0$, there exists a $\overline{k} \in \mathbb{N}$ such that $h_{k_j} < \frac{2\epsilon}{N}$ for all $k_j \ge \overline{k}$. Consequently, due to equation (81),

$$\left|\tilde{\mathrm{Tr}}(H^j\gamma_{k_j}) - \mathrm{Tr}(H^j\gamma_{k_j})\right| < \left|\frac{\epsilon}{N}\sum_{q=1}^{k_j} c_q^{k_j} n_q^{k_j}\right| < \epsilon.$$

This concludes the proof of (79). □

After the convergence of $\tilde{\mathrm{Tr}}(\cdot)$ to $\mathrm{Tr}(\cdot)$ has been established, we are now ready to prove the announced Γ-convergence result.

Lemma 6.6 *For every $\phi_j \rightharpoonup \phi$ in $\mathcal{V}$, every $u_j \rightharpoonup u$ in $\mathcal{U}$ and all $\gamma \in \mathcal{X}$,*

$$\tilde{\mathrm{Tr}}\big(H^j(\phi_j,u_j)\gamma\big) + I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma) \xrightarrow{\Gamma} \mathrm{Tr}\big(H(\phi,u)\gamma\big) + I_{\mathcal{K}_N^{H(\phi,u)}}(\gamma)$$

in the limit $j \to \infty$.

Proof Let us begin with the liminf part of the Γ-convergence proof. From Lemma 6.1, we have that for all $\phi_j \rightharpoonup \phi$ in $\mathcal{V}$ and all $u_j \rightharpoonup u$ in $\mathcal{U}$, for every $\gamma \in \mathcal{X}$ and all $\gamma_j \overset{*}{\rightharpoonup} \gamma$,

$$\mathrm{Tr}\big(H(\phi,u)\gamma\big) + I_{\mathcal{K}_N^{H(\phi,u)}}(\gamma) \le \liminf_{j\to\infty}\Big\{\mathrm{Tr}\big(H^j(\phi_j,u_j)\gamma_j\big) + I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma_j)\Big\}.$$

Using Lemma 6.5,

$$\begin{aligned}
&\liminf_{j\to\infty}\Big\{\mathrm{Tr}\big(H(\phi,u)\gamma\big)+I_{\mathcal{K}_N^{H(\phi,u)}}(\gamma)\Big\}\\
&\quad\le\liminf_{j\to\infty}\Big\{\tilde{\mathrm{Tr}}\big(H^j(\phi_j,u_j)\gamma_j\big)-\mathrm{Tr}\big(H^j(\phi_j,u_j)\gamma_j\big)\Big\}+\liminf_{j\to\infty}\Big\{\mathrm{Tr}\big(H^j(\phi_j,u_j)\gamma_j\big)+I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma_j)\Big\}\\
&\quad\le\liminf_{j\to\infty}\Big\{\tilde{\mathrm{Tr}}\big(H^j(\phi_j,u_j)\gamma_j\big)-\mathrm{Tr}\big(H^j(\phi_j,u_j)\gamma_j\big)+\mathrm{Tr}\big(H^j(\phi_j,u_j)\gamma_j\big)+I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma_j)\Big\}\\
&\qquad=\liminf_{j\to\infty}\Big\{\tilde{\mathrm{Tr}}\big(H^j(\phi_j,u_j)\gamma_j\big)+I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma_j)\Big\}.
\end{aligned}$$

Similarly, for the limsup part, using the same recovery sequence $\{\gamma_{k_j}\}_{j\in\mathbb{N}}$ as the one constructed in Lemma 6.2,

$$\begin{aligned}
&\limsup_{j\to\infty}\Big\{\tilde{\mathrm{Tr}}\big(H^j(\phi_j,u_j)\gamma_{k_j}\big)+I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma_{k_j})\Big\}\\
&=\limsup_{j\to\infty}\Big\{\tilde{\mathrm{Tr}}\big(H^j(\phi_j,u_j)\gamma_{k_j}\big)-\mathrm{Tr}\big(H^j(\phi_j,u_j)\gamma_{k_j}\big)+\mathrm{Tr}\big(H^j(\phi_j,u_j)\gamma_{k_j}\big)+I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma_{k_j})\Big\}\\
&\le\limsup_{j\to\infty}\Big\{\tilde{\mathrm{Tr}}\big(H^j(\phi_j,u_j)\gamma_{k_j}\big)-\mathrm{Tr}\big(H^j(\phi_j,u_j)\gamma_{k_j}\big)\Big\}+\limsup_{j\to\infty}\Big\{\mathrm{Tr}\big(H^j(\phi_j,u_j)\gamma_{k_j}\big)+I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma_{k_j})\Big\}\\
&\le\limsup_{j\to\infty}\Big\{\mathrm{Tr}\big(H(\phi,u)\gamma\big)+I_{\mathcal{K}_N^{H(\phi,u)}}(\gamma)\Big\}.
\end{aligned}$$

Therefore, using the results of Lemma 6.2,

$$\limsup_{j\to\infty}\Big\{\tilde{\mathrm{Tr}}\big(H^j(\phi_j,u_j)\gamma_{k_j}\big)+I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma_{k_j})\Big\}\le\mathrm{Tr}\big(H(\phi,u)\gamma\big)+I_{\mathcal{K}_N^{H(\phi,u)}}(\gamma).$$

This completes the proof. □

Lemma 6.7 *If $u_j \rightharpoonup u$ in $\mathcal{U}$ and $\phi_j \rightharpoonup \phi$ in $\mathcal{V}$, then for every $\gamma\in\mathcal{X}$, the family of functionals $\big\{\tilde{\mathrm{Tr}}\big(H^j(\phi_j,u_j)\gamma\big)+I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma)\big\}_{j\in\mathbb{N}}$ is equi-coercive.*

Proof From Lemma 6.5, we have for every $\gamma\in\mathcal{K}_{N,j,k_j}^{H^j(\phi_j,u_j)}$

$$\begin{aligned}
\tilde{\mathrm{Tr}}\big(H^j(\phi_j,u_j)\gamma\big)-\mathrm{Tr}\big(H^j(\phi_j,u_j)\gamma\big)=&\sum_{q=1}^{k_j}\sum_{i=1}^{\infty}(m_q^{k_j}-\nu_q^{k_j})c_q^{k_j}\big(\mu_{e_i,e_i}(t_{q+1}^{k_j})-\mu_{e_i,e_i}(t_q^{k_j})\big)\\
\ge&\sum_{q=1}^{k_j}\sum_{i=1}^{\infty}(\lambda_{\mathrm{L}B}-\lambda_{\mathrm{UB}})c_q^{k_j}\big(\mu_{e_i,e_i}(t_{q+1}^{k_j})-\mu_{e_i,e_i}(t_q^{k_j})\big)\\
\ge&(\lambda_{\mathrm{L}B}-\lambda_{\mathrm{UB}})N,
\end{aligned}$$

where $(\lambda_{\mathrm{L}B},\lambda_{\mathrm{UB}})$ denote the a-priori given bounds on the spectrum of $H(\phi,u)$ for the binning algorithm.

Hence from Lemma 6.3, especially equation (78),

$$\begin{aligned}\tilde{\mathrm{Tr}}\big(H^j(\phi_j,u_j)\gamma\big)+I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma) &= \tilde{\mathrm{Tr}}\big(H^j(\phi_j,u_j)\gamma\big)-\mathrm{Tr}\big(H^j(\phi_j,u_j)\gamma\big)\\ &\quad+\mathrm{Tr}\big(H^j(\phi_j,u_j)\gamma\big)+I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma)\\ &\geq \frac{1}{2}\mathrm{Tr}(-\Delta\gamma)-C_{12}\|\sqrt{\rho_\gamma}\|^{\frac{3}{2}}_{L^2(\Omega)}+(\lambda_{\mathrm{L}B}-\lambda_{\mathrm{U}B})N.\end{aligned}$$

This shows that for any $t\in\mathbb{R}$ the level sets

$$\Big\{\gamma\in\mathcal{X}:\tilde{\mathrm{Tr}}\big(H^j(\phi_j,u_j)\gamma\big)+I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma)<t\Big\},$$

are bounded,

$$t\geq\frac{1}{2}\|\gamma\|_{\mathcal{X}}-C_{12}\|\sqrt{\rho_\gamma}\|^{\frac{3}{2}}_{\mathcal{L}^2(\Omega)}-\frac{N}{2}+(\lambda_{\mathrm{LB}}-\lambda_{\mathrm{U}B})N.\qquad\square$$

Lemma 6.8 *If $\phi_j\rightharpoonup\phi$ in $\mathcal{V}$ and $u_j\rightharpoonup u$ in $\mathcal{U}$, then*

$$\lim_{j\to\infty}\inf_{\gamma\in\mathcal{X}}\Big\{\tilde{\mathrm{Tr}}(H^j(\phi_j,u_j)\gamma)+I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma)\Big\}=\inf_{\gamma\in\mathcal{X}}\Big\{\mathrm{Tr}\big(H(\phi,u)\gamma\big)+I_{\mathcal{K}_N^{H(\phi,u)}}(\gamma)\Big\}.$$

Proof This is a direct consequence of Theorem 7.8 in [12], Lemma 6.6 and Lemma 6.7. □

6.3 Γ-convergence of the operators S^{j,k_j}

In the next step we consider the Γ-convergence of $-S^{j,k_j}(u_j,\phi)$ to $-S(u,\phi)$ for $u_j\rightharpoonup u$.

Lemma 6.9 *If $u_j\rightharpoonup u$ in $\mathcal{U}$, then for $j\to\infty$,*

$$-S^{j,k_j}(u_j,\phi)\xrightarrow{\Gamma}-S(u,\phi)$$

with respect to the weak topology in $\mathcal{V}$.

Proof From Lemma 6.8, for every $u\in\mathcal{U}$ and all $u_j\rightharpoonup u$ in $\mathcal{U}$,

$$\lim_{j\to\infty}\inf_{\gamma\in\mathcal{X}}\Big\{E_{\mathrm{band}_{j,k_j}}(u_j,\phi,\gamma)+I_{\mathcal{K}_{N,k_j}^{H^j(\phi,u_j)}}(\gamma)\Big\}=\inf_{\gamma\in\mathcal{X}}\Big\{E_{\mathrm{band}}(u,\phi,\gamma)+I_{\mathcal{K}_N^{H(\phi,u)}}(\gamma)\Big\}.$$

Beginning with the liminf condition, for every $\phi\in\mathcal{V}$ and all $\phi_j\rightharpoonup\phi$ in $\mathcal{V}$,

$$\int_\Omega C_S|\nabla\phi(\mathbf{r})|^2\,\mathrm{d}\mathbf{r}\leq\liminf_{j\to\infty}\int_\Omega C_S|\nabla\phi_j(\mathbf{r})|^2\,\mathrm{d}\mathbf{r},$$

and

$$-\int_\Omega b(\mathbf{r},\{\mathbf{R}\})\phi(\mathbf{r})\,\mathrm{d}\mathbf{r}\leq\liminf_{j\to\infty}\bigg(-\int_\Omega b(\mathbf{r},\{\mathbf{R}\})\phi_j(\mathbf{r})\,\mathrm{d}\mathbf{r}\bigg).$$

This shows

$$-S(u,\phi)\leq\liminf_{j\to\infty}\big(-S^{j,k_j}(u_j,\phi)\big).$$

For the limsup condition, we can pick the recovery sequence $\tilde{\phi}_j$ to be the projection of $\phi \in \mathcal{V}$ onto $\mathcal{V}_j$. From the density of the spaces $\mathcal{V}_j$ as $j \to \infty$, we have $\tilde{\phi}_j \to \phi$ in $\mathcal{V}$. Hence, for this recovery sequence, we obtain

$$\lim_{j\to\infty} \int_\Omega C_S |\nabla \tilde{\phi}(\mathbf{r})|^2 \,\mathrm{d}\mathbf{r} = \int_\Omega C_S |\nabla \phi(\mathbf{r})|^2 \,\mathrm{d}\mathbf{r}$$

and

$$\lim_{j\to\infty} \left(-\int_\Omega b(\mathbf{r},\{\mathbf{R}\}) \tilde{\phi}_j(\mathbf{r}) \,\mathrm{d}\mathbf{r} \right) = -\int_\Omega b(\mathbf{r},\{\mathbf{R}\}) \phi(\mathbf{r}) \,\mathrm{d}\mathbf{r}.$$

In conclusion, for $u_j \rightharpoonup u$, the Γ-convergence of $-S^{j,k_j}(u_j,\phi)$ to $-S(u,\phi)$ has been established. □

Lemma 6.10 *If $u_j \rightharpoonup u$ in $\mathcal{U}$, then the family of functionals $\{-S^{j,k_j}(u_j,\phi)\}_{j\in\mathbb{N}}$ is equi-coercive with respect to the weak topology in $\mathcal{V}$.*

Proof Proceeding as in Lemma 6.3, we find

$$\begin{aligned} -S^{j,k_j}(u_j,\phi) \;=& \int_\Omega \big(C_S |\nabla\phi(\mathbf{r})|^2 - b(\mathbf{r},\{\mathbf{R}\})\phi(\mathbf{r}) \big) \,\mathrm{d}\mathbf{r} \\ & - \inf_{\gamma\in\mathcal{X}} \Big\{ \tilde{\mathrm{Tr}}\big(H^j(\phi,u_j)\gamma\big) + I_{\mathcal{K}_{N,k_j}^{H^j(\phi,u_j)}}(\gamma) \Big\} + I_{\mathcal{V}_j}(\phi) \\ \ge& \, C_S \|\nabla\phi\|^2_{\mathcal{L}^2(\Omega)} - \|b(\mathbf{r},\{\mathbf{R}\})\|_{\mathcal{L}^2(\Omega)} \|\phi\|_{\mathcal{L}^2(\Omega)} - \mathrm{Tr}\big(H^j(\phi,u_j)\hat{\gamma}_j\big) + \epsilon_{k_j}. \end{aligned} \tag{82}$$

Here, $\hat{\gamma}_j \in \mathcal{K}_{N,k_j}^{H^j(\phi,u_j)}$ are minimal in (82) and satisfy for all $j \in \mathbb{N}$

$$\tilde{\mathrm{Tr}}\big(H^j(\phi,u_j)\hat{\gamma}_j\big) = \mathrm{Tr}\big(H^j(\phi,u_j)\hat{\gamma}_j\big) - \epsilon_{k_j},$$

where due to Lemma 6.5 the sequence ϵ_{k_j} converges to 0 as j becomes infinite. It follows

$$\begin{aligned} -S^{j,k_j}(u_j,\phi) \;\ge& \, C_{13}\|\phi\|^2_{\mathcal{L}^2(\Omega)} - (\|b(\mathbf{r},\{\mathbf{R}\})\|_{\mathcal{L}^2(\Omega)} + \|\rho_{\hat{\gamma}_j}\|_{\mathcal{L}^2(\Omega)})\|\phi\|_{\mathcal{L}^2(\Omega)} \\ & - \|u_j\|_{\mathcal{L}^2(\Omega)}\|\rho_{\hat{\gamma}_j}\|_{\mathcal{L}^2(\Omega)} + \frac{1}{2}\mathrm{Tr}(-\Delta\hat{\gamma}) + \epsilon_{k_j} \\ \ge& \, C_{13}\|\phi\|^2_{\mathcal{L}^2(\Omega)} - C_{14}\|\phi\|_{\mathcal{L}^2(\Omega)} + C_{15}, \end{aligned} \tag{83}$$

with a constant $C_{13} > 0$ originating from the Poincaré inequality, and with further constants

$$\begin{aligned} C_{14} &\equiv \|b(\mathbf{r},\{\mathbf{R}\})\|_{\mathcal{L}^2(\Omega)} + \sup_{j\in\mathbb{N}} \|\rho_{\hat{\gamma}_j}\|_{\mathcal{L}^2(\Omega)}, \\ C_{15} &\equiv \sup_{j\in\mathbb{N}} \Big\{ -\|u_j\|_{\mathcal{L}^2(\Omega)}\|\rho_{\hat{\gamma}_j}\|_{\mathcal{L}^2(\Omega)} + \frac{1}{2}\mathrm{Tr}(-\Delta\hat{\gamma}_j) + \epsilon_{k_j} \Big\}. \end{aligned}$$

With (83), the equi-coercivity of $-S^{j,k_j}(u_j,\phi)$ with respect to the weak topology in $\mathcal{V}$ is proved. □

Lemma 6.11 *If $u_j \rightharpoonup u$ in $\mathcal{U}$, then $\lim_{j\to\infty} \sup_{\phi\in\mathcal{V}} S^{j,k_j}(u_j,\phi) = \sup_{\phi\in\mathcal{V}} S(u,\phi)$.*

Proof This is proven using Theorem 7.8 in [12], Lemma 6.9 and Lemma 6.10. □

6.4 Γ-convergence of the operators T^{j,k_j}

Lemma 6.12 *The family of functionals $\{T^{j,k_j}(u)\}_{j\in\mathbb{N}}$ converges in the* Γ*-sense, i.e. for $j \to \infty$*

$$T^{j,k_j}(u) \xrightarrow{\Gamma} T(u)$$

with respect to the weak topology in $\mathcal{U}$.

Proof We begin by showing the lim-inf condition for

$$T^{j,k_j}(u) = B^*_{\text{xc}}(u) + \sup_{\phi\in\mathcal{V}} S^{j,k_j}(u,\phi).$$

From Lemma 6.11, we have for every $u_j \rightharpoonup u$ in $\mathcal{U}$ and $u \in \mathcal{U}$,

$$\lim_{j\to\infty} \sup_{\phi\in\mathcal{V}} S^{j,k_j}(u_j,\phi) = \sup_{\phi\in\mathcal{V}} S(u,\phi).$$

In addition, $B^*_{\text{xc}}(u)$ is weakly lower semi-continuous, see [13]. Hence the liminf condition is proved.

In order to prove the limsup condition, for every $u \in \mathcal{U}$, let the recovery sequence $\{u_j\}_{j\in\mathbb{N}}$ be the projections of u onto $\mathcal{U}_j$. For this recovery sequence, using the bounds from equation (89) in the appendix B, the continuity of the functional $B^*_{\text{xc}}(u)$ in $\mathcal{U}$ can be established through Fatou's Lemma,

$$\lim_{j\to\infty} B^*_{\text{xc}}(u_j) = B^*_{\text{xc}}(u).$$

Hence, we have satisfied the limsup condition and have proven that in the limit $j \to \infty$, the family of functionals $T^{j,k_j}(u)$ converges in the Γ-sense with respect to the weak topology of $\mathcal{U}$ to $T(u)$. □

Lemma 6.13 *The family of functionals $\{T^{j,k_j}(u)\}_{j\in\mathbb{N}}$ is equi-coercive with respect to the weak topology in $\mathcal{U}$.*

Proof From Proposition 1.2 in [13],

$$B^*_{\text{xc}}(u) = \int_\Omega h^*\big(u(\mathbf{r})\big)\,\mathrm{d}\mathbf{r},$$

where $h^*(x) : \mathbb{R} \to \mathbb{R}$ is the Legendre transform of $(-h(t))$ from equation (10). Using the bounds from equation (89) in Appendix B, there exist real constants $C_{16} > 0$ and C_{17} such that

$$B^*_{\text{xc}}(u) \geq C_{16}\|u\|^4_{\mathcal{U}} - C_{17}(\text{vol}\Omega). \tag{84}$$

The estimate (84) implies natural bounds from below on the functional T^{j,k_j},

$$\begin{aligned}
T^{j,k_j}(u) &= B^*_{\text{xc}}(u) + \sup_{\phi\in\mathcal{V}} S^{j,k_j}(u,\phi) \\
&\geq B^*_{\text{xc}}(u) + \inf_{\gamma\in\mathcal{X}} \Big\{ \tilde{\text{Tr}}\big(H^j(\hat{\phi},u)\gamma\big) + I_{K^{H^j(\hat{\phi},u)}_{N,k_j}}(\gamma) \Big\} \\
&\geq B^*_{\text{xc}}(u) + N\lambda_{\text{LB}}(\hat{\phi},u) \\
&\geq B^*_{\text{xc}}(u) + N\Big(\lambda_1^{H^j(\hat{\phi},u)} + C_j\Big),
\end{aligned}$$

where $\hat{\phi} = 0$ is a test function in $\mathcal{V}$, λ_{LB} denotes the lower bound of the binning interval $[\lambda_{\text{LB}}, \lambda_{UB}]$ for $H^j(\hat{\phi}, u)$, and $\lambda_1^{H^j(\hat{\phi},u)}$ denotes the lowest eigenvalue of $H^j(\hat{\phi}, u)$. Let

$$\lambda_{\text{LB}} = \lambda_1^{H^j(\hat{\phi},u)} + C_j.$$

We know that $\sup_j |C_j|$ is uniformly bounded, because λ_{LB} is only a functional of $\hat{\phi}$ and u and independent of spatial discretization.

If $\xi_1^{H^j(\hat{\phi},u)}$ denotes the corresponding normalized eigenvector of $H^j(\hat{\phi}, u)$, we can derive a lower bound of $\lambda_1^{H^j(\hat{\phi},u)}$ by the ellipticity of the underlying variational problem,

$$\begin{aligned}\lambda_1^{H^j(\hat{\phi},u)} &= \left\langle H^j(\hat{\phi}, u)\xi_1^{H^j(\hat{\phi},u)}, \xi_1^{H^j(\hat{\phi},u)} \right\rangle \\ &\geq \|\nabla \xi_1^{H^j(\hat{\phi},u)}\|^2_{\mathcal{L}^2(\Omega)} - \|u\|_{\mathcal{L}^2(\Omega)} \\ &\geq -\|u\|_{\mathcal{L}^2(\Omega)}. \end{aligned} \tag{85}$$

Using the inequality (85), we can bound $T^{j,k_j}(u)$ from below by a coercive functional which is independent of j and k_j,

$$\begin{aligned} T^{j,k_j}(u) &\geq B^*_{\text{xc}}(u) - N\|\hat{\phi} - u\|_{\mathcal{L}^2(\Omega)} \\ &\geq C_{16}\|u\|^4_{\mathcal{U}} - N\|u\|^2_{\mathcal{U}}. \end{aligned} \tag{86}$$

In the limit $\|u\|_{\mathcal{U}} \to \infty$, the term $C_{16}\|u\|^4_{\mathcal{U}}$ dominates, so we have $T^{j,k_j}(u) \to \infty$. Hence, the equi-coercivity of the family of functionals $T^{j,k_j}(u)$ is established. □

Theorem 3 *In the limit of the number of spatial discretizations $j \to \infty$, and consequently in the limit of the number of spectral discretizations $k_j \to \infty$, the family of ground-state energies of the spatially and spectrally discrete K-S energy functionals converges to the full K-S ground-state energy,*

$$\lim_{j\to\infty} \inf_{u\in\mathcal{U}} T^{j,k_j}(u) = \inf_{u\in\mathcal{U}} T(u) = \varepsilon_{\text{GS}}.$$

Alternatively, in terms of the functional $L(u, \phi, \gamma)$, this means

$$\lim_{j\to\infty} \inf_{\mathcal{U}_j} \sup_{\mathcal{V}_j} \inf_{\mathcal{K}^{H^j(\phi,u)}_{N,k_j}} L(u, \phi, \gamma) = \inf_{\mathcal{U}} \sup_{\mathcal{V}} \inf_{\mathcal{K}^{H(\phi,u)}_N} L(u, \phi, \gamma) = \varepsilon^{\text{REKS}}_{\text{GS}}.$$

Proof This is proven using Theorem 7.8 in [12], Lemma 6.12 and Lemma 6.13. □

7 Binning in one dimension, a model problem

We now test the efficiency of the binning algorithm on a one-dimensional model problem which was first proposed in [11].

Consider a linear chain of M atoms with N electrons spaced uniformly with $R_i = i$ for $i \in \mathbb{Z}$. The electrons in the atoms are non-interacting electrons that interact with an effective field that depends on the positions of the nuclei in the chain. The effective potential $V(r)$ is a sum of Gaussian potentials centered at each atom in the chain,

$$V(r) = -\sum_{i\in\mathbb{Z}} \frac{\alpha}{\sqrt{2\pi\beta^2}} \exp\left[\frac{-(r - R_i)^2}{2\beta^2}\right].$$

Finding the ground-state energy of the system amounts to finding the N lowest eigenvalues of the linear eigenvalue problem in one dimension,

$$H\psi_i = \Big(-\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}r^2} + V(r) \Big)\psi_i = \epsilon_i \psi_i.$$

The constants α and β in the effective potential dictate the band gap in the band-structure of the one-dimensional chain. Hence the model has the ability to simulate either a metal or an insulator. In this paper, we test the binning algorithm on a "metallic" chain (setting $\alpha = 10$, $\beta = 0.45$), and an "insulating" chain (setting $\alpha = 100$, $\beta = 0.3$).

7.1 Binning algorithm for a linear eigenvalue problem

The binning algorithm works as follows:

```
do Find an initial guess to $[\lambda_{LB}, \lambda_{UB}]$;
Perform a $LDL^T$ decomposition of $H^j - \lambda_{LB}\mathcal{I}^j$ and $H^j - \lambda_{UB}\mathcal{I}^j$;
Find $\mathcal{N}_-(H^j - \lambda_{LB}\mathcal{I}^j)$ and $\mathcal{N}_-(H^j - \lambda_{UB}\mathcal{I}^j)$;
if $\mathcal{N}_-(H^j - \lambda_{LB}\mathcal{I}^j) > 0$;
then
    Decrease $\lambda_{LB}$ until $\mathcal{N}_-(H^j - \lambda_{LB}\mathcal{I}^j) = 0$.
end
if $\mathcal{N}_-(H^j - \lambda_{UB}\mathcal{I}^j) < N$;
then
    Increase $\lambda_{UB}$ until $\mathcal{N}_-(H^j - \lambda_{UB}) > N$;
else
    Use bisection to decrease $\lambda_{UB}$ so that $\mathcal{N}_-(H^j - \lambda_{UB}\mathcal{I}^j) = N + \epsilon_N$ with $\epsilon_N \in \mathbb{N}_{>0}$;
end
do Partition $[\lambda_{LB}, \lambda_{UB}]$ into $k$ intervals with end points $\{t_0^k, t_1^k, \dots, t_k^k\}$, $\lambda_{LB} = t_0^k$ and
$\lambda_{UB} = t_k^k$;
for q=1:k;
do
    Perform a $LDL^T$ decomposition of $H^j - t_q^k\mathcal{I}^j$ and find $\mathcal{N}_-(H^j - t_q^k\mathcal{I}^j)$;
end
for q=1:k;
do
    $n_q^{k,j} = \mathcal{N}_-(H^j - t_q^k\mathcal{I}^j) - \mathcal{N}_-(H^j - t_{q-1}^k\mathcal{I}^j)$;
    $m_q^k = \frac{(t_q^k + t_{q-1}^k)}{2}$;
end
do Minimize $\sum_{q=1}^{k} c_q^k m_q^k n_q^{k,j}$ over coefficients $\{c_q^k\} \subset \mathbb{R}^k$ subject to the constraints
$0 \le c - q^k \le 1$ and $\sum_{q=1}^{k} c_q^k n_q^{k,j} = N$.
```

Algorithm 1: Binning Algorithm

A system of 1000 atoms and 4000 electrons with periodic boundary conditions is discretized using a 8-th order central difference stencil in finite difference. To find an initial guess of $[\lambda_{LB}, \lambda_{UB}]$, we use the smallest and largest Ritz values obtained from a Krylov subspace projection of dimension k on an arbitrary unit vector, where k denotes the number of bins. Note that any Krylov subspace with dimension $p \geq 2$ may be used to obtain an initial guess of $[\lambda_{\mathrm{L}B}, \lambda_{\mathrm{U}B}]$. We use the interior-point method to perform the minimization of (53) with respect to the spectral binning coefficients $\{c_q^k\}_{q=1}^k$ subject to the constraints in equation (54).

The band energy of a "metallic" and an "insulating" system have been calculated using spectral binning and linear-scaling spectral Gauss quadratures (LSSGQ) from [27]. We see that spectral binning can achieve comparable accuracies as polynomial approximations.

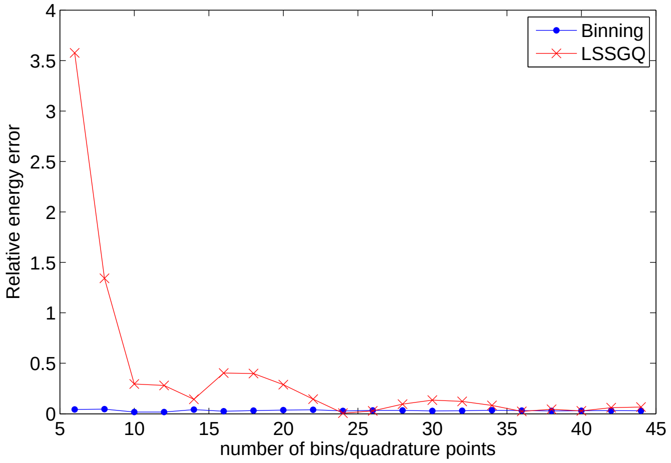

Figure 1: Metal: $\alpha = 10$, $\beta = 0.45$

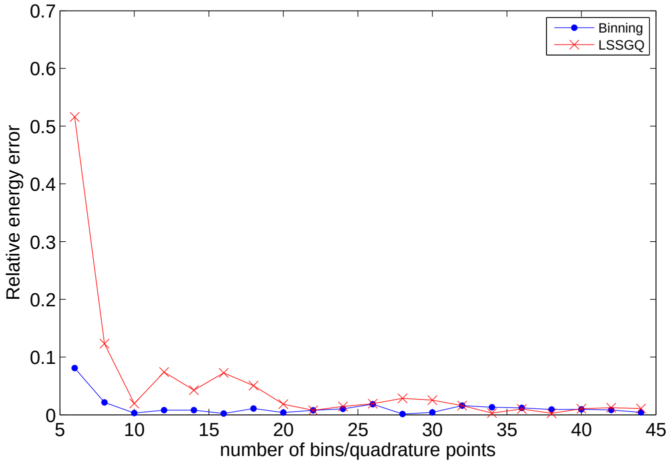

Figure 2: Insulator: $\alpha = 100$, $\beta = 0.3$

7.2 Electron density $\rho(\mathbf{r})$ with spectral binning

The electron density $\rho(\mathbf{r})$ in the context of spectral binning is

$$\begin{aligned}\rho_\gamma(\mathbf{r}_0) = \gamma(\mathbf{r}_0, \mathbf{r}_0) = \langle \mathbf{r}_0, \gamma \mathbf{r}_0 \rangle = \sum_{q=1}^{k} c_q^k \langle \mathbf{r}_0, s_{t_q^k}(H) \mathbf{r}_0 \rangle \\ = \sum_{q=1}^{k} c_q^k \sum_{p=1}^{\infty} s_{t_q^k}(\lambda_p) \langle \mathbf{r}_0, \xi_p \rangle \langle \xi_p, \mathbf{r}_0 \rangle = \sum_{q=1}^{k} c_q^k \sum_{p=1}^{\infty} s_{t_q^k}(\lambda_p) |\xi_p(\mathbf{r}_0)|^2 \\ = \sum_{q=1}^{k} c_q^k \sum_{p=1}^{\infty} s_{t_q^k}(\lambda_p) \left| \sum_{m=1}^{\infty} b_m^p e_m(\mathbf{r}_0) \right|^2 ,\end{aligned} \tag{87}$$

where

$$b_m^p \equiv \langle \xi_p, e_m \rangle, \qquad s_{t_q^k}(\lambda_p) \equiv \begin{cases} 1, & \text{if } t_q^k \leq \lambda_p \leq t_{q+1}^k, \\ 0, & \text{otherwise} \end{cases}$$

for an orthonormal basis set $\{e_m\}_{m \in \mathbb{N}}$, and the eigen-pairs of H are denoted by $\{\lambda_p, \xi_p\}$. In the form of a spectral integral, as shown in [27], equation (87) can be written as

$$\sum_{p=1}^{\infty} s_{t_q^k}(\lambda_p) \left| \sum_{m=1}^{\infty} b_m^p e_m(\mathbf{r}_0) \right|^2 = \int_{\sigma(H)} s_{t_q^k}(\lambda) \, \mathrm{d}\mu_{(\eta_{\mathbf{r}_0}, \eta_{\mathbf{r}_0})}$$

and

$$\rho(\mathbf{r}_0) = \sum_{q=1}^{k} c_q^k \int_{\sigma(H)} s_{t_q^k}(\lambda) \, \mathrm{d}\mu_{(\eta_{\mathbf{r}_0}, \eta_{\mathbf{r}_0})},$$

where

$$\eta_{\mathbf{r}_0}(\mathbf{r}) = \sum_{p=1}^{\infty} e_p(\mathbf{r}_0) e_p(\mathbf{r}).$$

In other words, evaluation of the electron density using spectral binning requires the ability to evaluate the quantity $\langle \eta_{\mathbf{r}_0}, s(H) \eta_{\mathbf{r}_0} \rangle$. An efficient approach to doing this without polynomial or rational approximations remains an open problem.

8 Acknowledgements

Part of this work was conducted when TB held a position at the Max-Planck-Insitute for Mathematics in the Sciences in Leipzig. We gratefully acknowledge the partial financial support of the US National Science Foundation under the PIRE Grant No. OISE-0967140. This provides partial support of XW, KB and MO, and enabled the visit of XW to the Max-Planck-Institute of Leipzig. The research was also partially sponsored by the US Army Research Laboratory and was partially accomplished under Cooperative Agreement Number W911NF-12-2-0022. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Appendix A Orbital formulation of K-S DFT

The K-S problem [22] constitutes the minimization of the functional

$$\int_{\Omega} \frac{1}{2} \sum_{1\leq i\leq N} |\nabla\psi_i|^2 \,\mathrm{d}\mathbf{r} + E_{\mathrm{H}}\left(\rho\right) + E_{\mathrm{ext}}\left(\rho\right) + E_{\mathrm{ZZ}} + E_{\mathrm{xc}}\left(\rho\right)$$

over

$$\left\{ \{\psi_i\} \in \mathcal{V}^N : \langle\psi_i, \psi_j\rangle = \delta_{ij} \right\},$$

where E_{H}, E_{ext}, E_{ZZ} and E_{xc} are given by (7), (8), (9) and (10), respectively, and with the electron density $\rho = \sum_{i=1}^{N} |\psi_i|^2$.

The Euler-Lagrange equation associated with the constrained variational problem above gives rise to the non-linear eigenvalue problem

$$\left(-\frac{1}{2}\Delta + V\right)\psi = \lambda\psi,$$

where

$$V\big(\rho(\mathbf{r}), \mathbf{r}\big) = \int_{\Omega} \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} \,\mathrm{d}\mathbf{r}' + \sum_{1\leq I\leq M} \frac{Z_I}{|\mathbf{R}_I - \mathbf{r}|} + h'(\rho(\mathbf{r})).$$

The solution to the variational problem is given by the eigenvectors ψ_i that correspond to the N lowest eigenvectors. The problem is non-linear because V depends on ρ and thus on ψ_i.

The operator formulation that we use is obtained formally by noting that any $\gamma \in \mathcal{K}_N$ has the representation

$$\gamma = \sum_{1\leq i\leq N} \psi_i \otimes \psi_i$$

for $\{\psi_i\} \subset \mathcal{V}^N$.

Appendix B The dual formulation of exchange-correlation

Let $\mathcal{T}$ be a topological vector space and $\{F_I\}$ be a family of continuous affine functionals from $\mathcal{T}$ to $\bar{\mathbb{R}}$. Let $\Gamma(\mathcal{T})$ denote the collection of functionals that are the point-wise supremum of some family $\{F_I\}$. Since the point-wise supremum of a family of convex functionals is convex and the point-wise supremum of a family of lower semi-continuous functionals is lower semi-continuous, see, e.g. [13], we have that every functional in $\Gamma(\mathcal{V})$ is convex and lower semi-continuous. Further, we have the following statement, see Proposition 3.1 in [13].

Proposition Appendix B.1 *The following properties are equivalent:*

1. *$F \in \Gamma(\mathcal{T})$.*
2. *F is a convex lower semi-continuous functional from $\mathcal{T}$ to $\bar{\mathbb{R}}$ and if F takes the value $-\infty$, then F is identically equal to $-\infty$.*

Given $F : \mathcal{T} \mapsto \bar{\mathbb{R}}$, the dual conjugate functional $F^* : \mathcal{T}^* \mapsto \bar{\mathbb{R}}$, where $\mathcal{T}^*$ denotes the space of linear functionals defined on $\mathcal{T}$, is

$$F^* = \sup_{u\in\mathcal{T}} \left\{ \langle u^*, u\rangle - F(u) \right\}.$$

We see that F^* is defined as the point-wise supremum of the family of continuous affine functionals $\langle \cdot, u\rangle - F(u)$, hence $F^* \in \Gamma(\mathcal{T}^*)$, and F^* is convex and lower semi-continuous. Furthermore, if F itself is convex and lower semi-continuous, the dual conjugate functional of F^* coincides with F, (i.e. $F^{**} = F$), see, e.g., Proposition 4.1 in [13].

When we apply the aforementioned properties of dual transforms to the exchange-correlation functional, since $-E_{\mathrm{xc}}(\rho_\gamma)$ is convex and lower semi-continuous in $\mathcal{L}^{\frac{4}{3}}(\Omega)$, we have $-E_{\mathrm{xc}}(\rho) \in \Gamma(\mathcal{L}^{\frac{4}{3}}(\Omega))$. We can then rewrite $-E_{\mathrm{xc}}(\rho)$ as

$$\begin{aligned} -E_{\mathrm{xc}}(\rho_\gamma) &= \sup_{u \in \mathcal{L}^{r'}(\Omega)} \{\langle u, \rho_\gamma\rangle - B_{\mathrm{xc}}(u)^*\} \\ &= - \inf_{u \in \mathcal{L}^{r'}(\Omega)} \{B_{\mathrm{xc}}^* - \langle u, \rho_\gamma\rangle\}, \end{aligned} \tag{88}$$

where

$$B_{\mathrm{xc}}^*(u) = (-E_{\mathrm{xc}}(\rho))^*$$

and B_{xc}^* is convex and lower semi-continuous in $\mathcal{L}^{r'}(\Omega)$ with $\frac{1}{r'} = 1 - \frac{1}{4/3} = \frac{1}{4}$. This also explains the choice of $\mathcal{U}$ in equation (16).

From Proposition 2.1 in [13], we know that

$$B_{\mathrm{xc}}^*(u) = \int_\Omega h^*(u)\,\mathrm{d}\mathbf{r},$$

where $h^*(x) = (-h(t))^* = \sup\limits_{t\in\mathbb{R}}\{xt - (-h(t))\}$ is the Legendre transform of the function $-h(t)$. Due to the bounds

$$C_1|t|^{\frac{4}{3}} + C_2 \le -h(t) \le C_3|t|^{\frac{4}{3}} + C_4$$

on $-h(t)$, we can arrive at the bounds

$$C_{18}|x|^4 + C_{19} \le h^*(x) \le C_{16}|x|^4 + C_{17} \tag{89}$$

for $h^*(x)$.

References

[1] B. Adams. *Sobolev Spaces. Academic Press, INC. 1978.*

[2] A. Anantharaman and E. Cancès. *Existence of minimizers for Kohn-Sham models in quantum chemistry. Annales de L'Institut Henri Poincare – Analyse Non-Lineaire*, 26(6):2425-2455, 2009.

[3] S. Ismail-Beigi and T. Arias. *New Algebraic Formulation of Density Functional Calculation, Computer Physics Communications*, 2000.

[4] W. Arveson. *Ten lectures on operator algebras. American Mathematical Society, 1980.*

[5] R. Baer and M. Head-Gordon. *Chebyshev expansion methods for electronic structure calculations on large molecular systems Journal of Chemical Physics, 107, 1997.*

[6] F.A. Berezin and M.A. Shubin. *The Schrödinger Equation. Kluwer Academic publishers, 1991.*

[7] D.R. Bolwer, T. Miyazaki. *Order-N methods in electronic structure calculations Reports on Progress in Physics* 75, 2009.

[8] A. Braides. *Gamma-Convergence for Beginners Oxford University Press, 2002.*

[9] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations. Springer, New York, 2010.*

[10] I. Catto, C. Le Bris and P.L. Lions. *On the thermodynamic limit for Hartree-Fock type models Annales de L'Institut Henri Poincare – Analyse Non-Lineaire*, 18(6): 687-760, 2001.

[11] E.J. García-Cervera, J. Liu, Y. Xuan, and W. E *Linear-scaling subspace iteration algorithm with optimally localized nonorthogonal wave functions for Kohn-Sham density functional theory Physical Review B*, 79, 2009.

[12] G. Dal Maso. *An Introduction to Gamma-Convergence Birkhäuser, 1993.*

[13] I. Ekeland and R. Temam. *Convex Analysis and Variational Problems. North-Holland, Amsterdam, 1976.*

[14] L.C. Evans. *Partial Differential Equations. American Mathematical Society, 2010.*

[15] J. Jost. *Calculus of Variations. Cambridge University Press, 1998.*

[16] V. Gavini, J. Knap, K. Bhattacharya, M. Ortiz. *Non-periodic finite-element formulation of orbital-free density functional theory. Journal of the Mechanics and Physics of Solids, 55:669-696, 2007.*

[17] M.J. Gillian. *Calculation of the vacancy formation energy in alumnium. Journal of Physics: condensed matter 1* 4, 1989.

[18] S. Goekecker and M. Teter. *Tight-binding electronic-structure calculations and tight-binding molecular dynamics with localized orbitals. Physical Review B* 51, 1995.

[19] S. Goekecker. *Linear scaling electronic structure methods. Reviews of Modern Physics* 71(4), 1999.

[20] L. Lin, M. Chen, C. Yang and Y. He. *Accelerating atomic orbital-based electronic structure calculation via pole expansion and selected inversion. Journal of Physics: condensed matter* 25(29)2013.

[21] B. Parlett. *The Symmetric Eigenvalue Problem. SIAM, 1998.*

[22] R. Parr and W. Yang. *Density-Functional Theory of Atoms and Molecules. Oxford University Press,1989.*

[23] J. Perdew and Y. Wang. *Accurate and simple analytic representation of the electron-gas correlation energy. Physical Review B* 45, 1992.

[24] M. Reed and B. Simon. *Functional Analysis. Academic Press, 1980.*

[25] W. Rudin. *Functional Analysis. McGraw-Hill, Boston, 1991.*

[26] C-K Skylaris, P. Haynes, A. Mostofi and M.C. Payne. *Introducing ONETEP: Linear-scaling density functional simulations on parallel computers Journal of Chemical Physics* 122, 2005.

[27] P. Suryanarayana, K. Bhattacharya and M. Ortiz. *Coarse graining Kohn-Sham density functional theory. Journals of Mechanics and Physic of Solids* 61(1) 3860, 2012.

[28] P. Suryanarayana, V. Gavini, T. Blesgen, K. Bhattacharya and M. Ortiz. *Non-periodic finite-element formulation of KohnSham density functional theory. Journals of Mechanics and Physic of Solids* 58 256-280, 2010.

[29] P. Suryanarayana. *On spectral quadrature for linear-scaling Density Functional Theory. Chemical Physics Letters* 584 182-187, 2013.

[30] P. Suryanarayana. *Optimized purification for density matrix calculation Chemical Physics Letters* 555 291-295, 2013.

[31] J.J. Sylvester. *A demonstration of the theorem that every homogeneous quadratic polynomial is reducible by real orthogonal substitutions to the form of a sum of positive and negative squares. Philosophical Magazine* 4 (23): 138-142, 1852.

[32] J. VandeVondele et al. *Quickstep: Fast and accurate density functional calculations using a mixed Gaussian and plane waves approach. Computer Physics Communications* 2(167), 2005.

[33] E.J.M. Veling. *Lower bounds for the infimum of the spectrum of the Schrodinger operator in $\mathbb{R}^N$ and the Sobolev inequalities. Journals of inequalities in pure and applied mathematics,* 3(4), 2002.

[34] R. Wheeden and A. Zygmund. *Measure and integral: an introduction to real analysis. Marcel Dekker, INC.*