A variational framework for spectral discretization of the density matrix in Kohn Sham density functional theory

Thesis by

Xin Wang

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2015

(Defended January 5, 2015)

To my parents, who left their home-country to give me a better education.

Acknowledgements

Getting a PhD is truly a humbling experience. I wouldn't have been able to make it this far without the unconditional love from my family, the support from my advisors and all of my friends here at Caltech. In the last five+ years, Caltech has become my second home.

I want to first thank my parents for their unwavering support in my education. My life would look very different today if they hadn't chosen to sacrifice their comfortable life in China to start a new and difficult life in this foreign country. Mom and dad, I am glad that I didn't disappoint you.

I have been very blessed to have two of the most knowledgable and supportive advisors at Caltech, Prof. Kaushik Bhattachaya and Prof. Michael Ortiz. From them, I learned what it means to be a scientist who hungers after knowledge; I learned how to persevere when confronted with seemingly unsolvable scientific questions. Thank you for giving me this opportunity to learn from you.

I had the honor to become friends with many loving and brilliant people through my study at Caltech. I want to thank my housemates, Megan Newcombe, Eyrun Eyjolfsdottir, and April Peet Vos for sharing their lives with me. Thank you for being there for me through all my ups-and-downs during my PhD. I want to thank Chirranjeevi (BG) Balaji Gopal, Steven Demers, Aron Varga, and Pratyush Tiwary for being great friends to whom I can always turn for company. You brought so much laughter into my life at Caltech. I am deeply grateful to have two wonderful friends, Jeff Amelang and Srivasan (Sri) Hulikal, for their generous help in my PhD research. Thank you for spending endless hours of your time discussing science with me. On this note, I would like to also thank friends from the dual group meeting,

Jonathan Chiang, Brandon Runnels and others. I have learned a great deal from you. I want to thank my collaborator, Prof. Thomas Blesgen for his meticulous contributions on our joint paper. I want to thank the class of 2009 PhD students from the Mechanical and Civil engineering, especially the wonderful ladies of our class, Marcella Gomez, Melissa Tanner, Swetha Veeraraghavan, and Jenny Jiang. Thank you for studying with me through the difficult first year and qualification exams. I want to thank the students and postdocs in Kaushik and Michael's group: Gal Schmuel, Zubaer Hossain, Mauricio Ponga, Likun Tan, Vinamra Agrawal, Lincoln Colins, Dingyi Sun, Paul Plucinsky, Chun-Jen Hseuh (Ren), Jin Yang, Paul Mazur, Landry Fakoua, Stephanie Heyden, Stephanie Mitchell and Sarah Mitchell, for insightful discussions throughout my PhD. I want to thank Stephanie Heyden and Aubrie Amelang for their friendship and all the delicious baked treats they've shared with me during my PhD.

Lastly and most importantly, I want to thank God for the joy while studying the laws of nature, as it is written in Psalm 111:2, "Great are the works of the Lord; They are studied by all who delight in them".

Contents

Nomenclature

α_I The nondimensional nucleus mass

$\mathbf{k}_f$ Th fermi level for momentum wavenumber

$\mathbf{R}_i$ The Ith nucleus spatial coordinate

$\mathbf{r}_i$ The ith electron spatial coordinate

$\bigwedge$ The space of antisymmetric functions

$\mathcal{D}_N$ The space of mixed-state N-electron density operators

$\mathcal{H}$ $\mathcal{L}(\Omega)$

$\mathcal{H}_e$ The space of anti-symmetric wavefunctions for N electrons

$\mathcal{H}_n$ The space of wavefunctions for M nuclei

χ_k^{OPW} Orthogonal plane-wave basis

$\mathcal{I}_{\{\}}()$ The indicator function for the set $\{\}$

$\mathcal{I}_N$ The space of orbitals for non-interacting electrons

$\mathcal{K}_N^j$ The spatially discretized space of $\mathcal{K}_N$

$\mathcal{K}_N^{H(\phi,u)}$ The vector space of density operators in $\mathcal{K}_N$ that commutes with the Hamiltonian $H(\phi, u)$

$\mathcal{K}_{N,k_j}^{H^j(\phi,u)}$ The vector space of density operators in $\mathcal{K}_N$ that commutes with the spatiall discretized Hamiltonian $H^j(\phi,u)$ in the span of the k binning basis $\{s_{t_q^k}\}_{q=1}^k$

$\mathcal{K}_N$ The vector space of self-adjoint, trace-class operator in $\mathcal{X}$ that has trace N

$\mathcal{U}$ The space of exchange-correlation potentials, $\mathcal{L}^4(\Omega)$

$\mathcal{U}_j$ The spatially discretized space of $\mathcal{U}$

$\mathcal{V}$ $\mathcal{W}_0^{1,2}(\Omega)$

$\mathcal{V}_j$ The spatially discretized space of $\mathcal{V}$

$\mathcal{X}$ The vector space of self-adjoint, trace-class operator on $\mathfrak{S}_1$ with finite kinetic energy

$\mathcal{X}_N$ The space of one-particle reduced density operators

$\Delta\mathbf{k}$ The volume per k-point

$\Delta\lambda$ The spectrum width of the Hamiltonian matrix

$\epsilon_0^e(\{\mathbf{R_1},\cdots,\mathbf{R_M}\})$ The ground state energy for electrons given nuclei positions at $\{\mathbf{R_1},\cdots,\mathbf{R_M}\}$

ϵ_0^{BO} The relaxed ground state energy of molecular system with Born-Oppenheimer approxmation

$\epsilon_0^{\mathrm{EKS}}$ The ground state energy for the extended Kohn-Sham functional

$\epsilon_0^{\mathrm{REKS}}$ The ground state energy for the reformulated extended Kohn-Sham functional

ϵ_0 The ground state energy of molecular system - with relaxation of the electrons and nuclei

$\epsilon_0^{\mathrm{KS}}(\{\mathbf{R_1},\cdots,\mathbf{R_M}\})$ The Kohn-Sham ground state energy

$\mathfrak{S}_1$ The vector space of self-adjoint, trace-class operator on $\mathcal{H}$

γ The one-particle reduced density operator

$\gamma(\mathbf{r}, \mathbf{r}')$ The one-particle reduced density operator in spatial coordinates

Γ_N The N-particle density operator

$\Gamma_{(N,\text{mixed})}$ The mixed N-particle density operator

$\hbar$ The reduced Planck's constant

λ^c Eigenvalue that correspond to core electrons

λ^v Eigenvalue that correspond to valence electrons

λ_f The fermi energy of the system

$\lambda_{\max}$ The lower bound of spectrum of Hamiltonian

$\lambda_{\min}$ The upper bound of spectrum of Hamiltonian

$\mathcal{K}_{n_p}$ The Krylov subspace of dimension n_p

$\mathcal{N}$ The space of electron densities that come from anti-symmetric wavefunctions with N electrons

$\mathcal{V}$ The space of ground state electron density (V-representable)

$\mathfrak{V}$ The V-representable density

μ The Lagrange multiplier for total number of electrons constraint

$\mu_{\xi,\xi}$ The spectral measure with respect to the vector ξ_i

$\phi(\mathbf{r})$ The electrostatic potential

$\Phi(\mathbf{r}, \mathbf{r}')$ The electrostatic potential operator

Ψ The many body wavefunction for N electrons and M nuclei

$\psi(\mathbf{r})$ The single electron orbital

$\psi^c(\mathbf{r})$ Eigenvector that correspond to core electrons

$\psi^v(\mathbf{r})$ Eigenvector that correspond to valence electrons

Ψ_e The manybody wavefunction for the electrons

Ψ_n The manybody wavefunctions for the nuclei

$\rho(\mathbf{r})$ The electron density

$\rho_\gamma(\mathbf{r})$ The electron density associated with the one-particle density operator γ

$\sigma(H^{\mathrm{KS}})$ The spectrum of H^{KS} operator

$\tilde{\psi}^v(\mathbf{r})$ Smoothed eigenvector that correspond to valence electrons

$\tilde{\mathrm{Tr}}$ The approximated Trace operator

$+, -$ The electron spin: up,down

ε_{c} The correlation integrand

ε_{x} The exchange integrand

$\{s_{t_q^k}\}_{q=1}^k$ The family of spectral binning basis

$\{t_q^k\}_{q=1}^k$ The collection of binning nodes

A_I The nucleus spin coordinate

$b(\mathbf{r}, \{\mathbf{R_1}, \cdots, \mathbf{R_M}\})$ The regularized nuclei charge density

$B_{\mathrm{xc}}^*(\rho)$ The dual functional for $B_{\mathrm{xc}}(\rho)$

$B_{\mathrm{xc}}(\rho)$ $-E_{\mathrm{xc}}(\rho)$, the negative of the exchange-correlation functional

C Computation cost

C_F Constant for homogeneous electron gas kinetic energy

$E^j_{\text{band}}(\gamma)$ The spectrally *discretized* band energy

$E^j_{\text{band}}(\gamma)$ The *discretized* band energy $\text{Tr}\big(H^j(\phi,u)\gamma\big)$

$E^{\text{EKS}}(\gamma)$ The extended Kohn-Sham energy functional

$E^{\text{REKS}}(\gamma)$ The reformulated extended Kohn-Sham energy functional

$E_{\text{band,j,k}_\text{j}}(\gamma)$ The spatially and spectrally *discretized* band energy

$E_{\text{band}}(u,\phi,\gamma)$ The band energy $\text{Tr}\big(H(\phi,u)\gamma\big)$

$E_{\text{c}}(\rho)$ The correlation energy functional

$E_{\text{H}}(\rho)$ Classical electrostatic repulsion energy functional, hartree energy

$E_{\text{xc}}(\gamma)$ The exchange-correlation functional written in terms of density operators γ

$E_{\text{xc}}(\rho)$ The exchange-correlation energy functional

$E_{\text{x}}(\rho)$ The exchange energy functional

$F_{LL}(\rho)$ The Levy-Lieb universal functional

$F_L(\rho)$ The Lieb universal functional

$g(\lambda)$ The matrix function for density operator

$g_{\text{fermi}}(\lambda)$ The Fermi-Dirac distribution

H The Hamiltonian Operator

$h(\rho)$ The exchange-correlation integrand

H^{KS} The Kohn-Sham Hamiltonian operator

H^{PS} Hamiltonian with pseudopotential

H_e The Hamiltonian operator for electrons

H_{box} The Hamiltonian operator for N identity non-interacting particles in the box

$J(\gamma)$ The Coulomb energy for the molecular system written as a function of the density operators γ

$J(\rho)$ The Coulomb energy for the molecular system

$L(\gamma)$ The Lagrangian functional for the reformulated Kohn-Sham energy functional

$L(R_l)$ The distance cutoff between localized basis centers for density matrix entry to be negligible

$L^j(\gamma)$ The *discretized* Lagrangian functional for the reformulated Kohn-Sham energy functional

M Number of nuclei in the molecular system

m^e The mass of an electron

m^n The mass of a nucleus

N Number of electrons

N_d Size of the Hamiltonian matrix

n_H The sparsity of Hamitlonian matrix

n_p Degree of Chebyshev polynomial approximation

n_x, n_y, n_z The quantum number in x,y,and z direction for particle in the box

$P(\lambda)$ The resolution of identity for H^{KS}

Q A partial isometry operator

q The charge of an electron

R_c The cut off radius for pseudopotentials.

R_l The localization region for Chebyshev approximations

$S(u, \phi)$ The functional for Columb energy

$S^{j,k_j}(u, \phi)$ The spatially and spectrally discretized functional for Column energy

T The kinetic energy of the homogeneous electron gas

$T(u)$ The dual functional for exchange-correlation

$T^{\mathrm{TF}}(\rho)$ The Thomas-Fermi kinetic energy functional

$T^{j,k_j}(u)$ The spatially and spectrally discretized dual functional for exchange-correlation

$T_0(\rho)$ Kinetic energy functional for non-interacting electrons

T_e The kinetic energy operator for interacting electrons

T_e The kinetic energy operator for the electrons

$T_e(\rho)$ Kinetic energy functional for interacting electrons

$T_J(\rho)$ The Janak kinetic energy functional

$T_k(x)$ Chebyshev polynomials

T_n The kinetic energy operator for the nuclei

T_V The kinetic energy per unit volume for homogeneous electron gas

$u(\mathbf{r})$ The exchange-correlation potential, dual function to electron density $\rho(\mathbf{r})$

$U(\mathbf{r}, \mathbf{r}')$ The exchange-correlation potential operator

U^{TF} The Thomas-Fermi interaction energies

U_{e-e} The electron-electron repulsion energy operator

$U_{e-e}(\rho)$ The electrostatic repulsion energy functional for interacting electrons

U_{n-e} The nucleus-electron attraction energy operator

U_{n-n} The nucleus-nucleus repulsion energy operator

V_{box} The potential for particles in the box

$v_{\text{ext}}(\mathbf{r})$ The external potential for the molecular system

$v_{\text{KS}}(\mathbf{r})$ The Kohn-Sham potential

$v_{\text{xc}}(\rho)$ The exchange-correlation potential

Z_I The charge of the Ith nucleus

$\mathbf{k}$ The electron momentum wave number for homogeneous electron gas

$\mathbf{n}$ The quantum number for particle in box

Abstract

Kohn-Sham density functional theory (KSDFT) is currently the main work-horse of quantum mechanical calculations in physics, chemistry, and materials science. From a mechanical engineering perspective, we are interested in studying the role of defects in the mechanical properties in materials. In real materials, defects are typically found at very small concentrations e.g., vacancies occur at parts per million, dislocation density in metals ranges from $10^{10}m^{-2}$ to $10^{15}m^{-2}$, and grain sizes vary from nanometers to micrometers in polycrystalline materials, etc. In order to model materials at realistic defect concentrations using DFT, we would need to work with system sizes beyond millions of atoms. Due to the cubic-scaling computational cost with respect to the number of atoms in conventional DFT implementations, such system sizes are unreachable. Since the early 1990s, there has been a huge interest in developing DFT implementations that have linear-scaling computational cost. A promising approach to achieving linear-scaling cost is to approximate the density matrix in KSDFT. The focus of this thesis is to provide a firm mathematical framework to study the convergence of these approximations. We reformulate the Kohn-Sham density functional theory as a nested variational problem in the density matrix, the electrostatic potential, and a field dual to the electron density. The corresponding functional is linear in the density matrix and thus amenable to spectral representation. Based on this reformulation, we introduce a new approximation scheme, called spectral binning, which does not require smoothing of the occupancy function and thus applies at arbitrarily low temperatures. We proof convergence of the approximate solutions with respect to spectral binning and with respect to an additional spatial discretization of the domain. For a standard one-dimensional benchmark

problem, we present numerical experiments for which spectral binning exhibits excellent convergence characteristics and outperforms other linear-scaling methods.

Chapter 1

Introduction

It is said that in experiments, we have a partial understanding of the full truth; and in computation, we have a full understanding of the partial truth. Therefore, in order to predict new material properties using computation, it is imperative that we build in as much physics as we can into the computational model, provided that it is still computationally feasible. Kohn-Sham Density functional theory (KSDFT) is precisely the theory for electron structure that strikes a good balance between minimizing empiricism in the model and maximizing computational efficiency.

Today, we find DFT in many applications: investigation of phase stability in various materials, oxides, thermoelectrics, ferroelectrics, e.g., Hautier *et al.* [27], Roy *et al.* [64], Doak and Wolverton [16], and Bennett *et al.* Bennett2011, etc; design of new alloys with superior structural properties, e.g., Sandlobes *et al.* [67], Trinkle *et al.* [77], and Hickel *et al.* [31], etc. More recently, DFT has become the primary tool for high throughput screening of materials, e.g., Saal *et al.* [66], Armiento *et al.* [4], etc.

The rapid increase in the number of publications involving DFT best illustrates the growing importance of DFT in physics, chemistry and materials science. Figure 1.1 plots the number of papers that contain the name "density functional theory" in their title and abstract from the web of science for the last 23 years. Unless there is another break-through in computational physics, we expect DFT to sustain its momentum for many years to come.

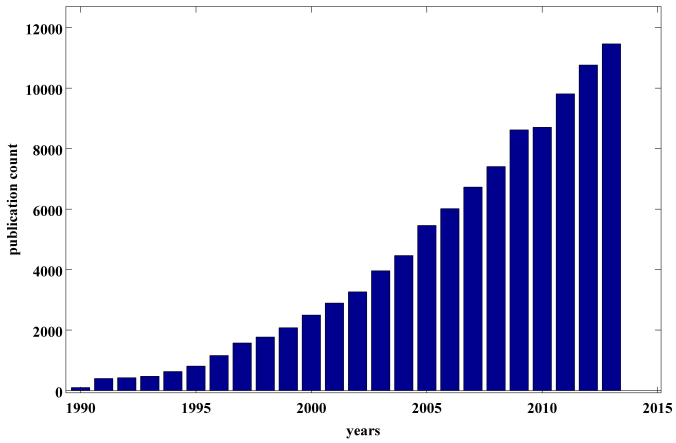

Figure 1.1: Number of publications with DFT as topic.

The development of DFT in 1964 by Walter Kohn was a huge break-through in physics because it linked the ground state energy of the molecular system to the ground state electron density. Kohn transformed the *linear* eigenvalue problem of finding the ground state of a molecular system from $3N$ dimensions to a *non-linear* eigenvalue problem in 3 dimensions, where N is the number of electrons in the systems. There are several good introductions to DFT. The two papers everyone who is interested in DFT should read are the pioneering papers written by Hohenberg and Kohn in 1964 [33] and Kohn and Sham in 1965 [38]. Other helpful introductions to DFT are by Parr and Yang [55], Martin [47], Cances[12], and Anantharaman and Cances [2].

The exchange correlation functional is crucial to the accuracy of a density functional theory calculation. The most basic exchange correlation functional, the local density approximation (LDA) was proposed by Kohn and Sham [38]. Widely used forms of LDA can be found in Perdew and Zunger [56] and Perdew and Wang [59]. A more sophisticate exchange correlation functional, the generalized gradient approximation (GGA) is introduced by Perdew [57]; different flavors of GGA exchange-correlation functionals can be found in [60], [40], [7], and [58]. Finally, there is the more recent development of hybrid functionals that mixes in the exact exchange energy [8].

For those who are interested in the development of pseudopotentials for density functional theory, the following papers would be useful: the earliest developments of pseudopotentials are found in Hellmann [28], Herring [30], Phillips and Kleinman [61], Antoncik [3]; on norm-conserving pseudopotential Hamann *et al.* [25], ultrasoft pseudopotential Vanderbilt [79], and separable pseudopotential operators Kleinman and Bylander [37] and Troullier and Martin [78]. Good review articles on pseudopotential can be found in Heine and Cohen [14], Harrison [26], and Pickett [62].

Lastly, for lower complexity algorithms in DFT, such as linear scaling methods, there has been numerous publications since the early 1990s. For density matrix expansion/approximation methods, there are Li *et al.* [42], Goedecker and Colombo [20], Hernandez *et al.* [29], Baer and Head-Gordon [6], Suryanarayana *et al.* [73], Lin *et al.* [45], Suryanarayana [71], Schofield *et al.* [68], and Nava *et al.* [52], etc, for methods that approximate the subspace spanned by the occupied orbitals, and there are Ordejon *et al.* [53], Mauri and Galli [50], Marzari and Vanderbilt [48], Garcia-Cervera *et al.* [18], and Motamarri and Gavini [51], etc. There are also two excellent review articles on linear scaling methods in DFT by Goedecker [23] and Bowler *et al.* [10].

Chapter 2

Density functional theory

I begin this introduction of DFT from the time-independent Schrödinger equation. Almost all of the information in this background section comes from the following two references, the first of which places more emphasis on explanation of the physics [55], and the second of which places more emphasis on mathematics [12].

2.1 Many-body Schrödinger equation

Consider an isolated molecule that consists of M nuclei and N electrons. The time-independent Schrödinger equation that governs the molecular system without accounting for relativistic effects is,

$$H\Psi = \epsilon\Psi, \tag{2.1}$$

where $\Psi : \mathbb{R}^{3(M+N)} \times \{+,-\} \to \mathbb{C}$ denotes the wavefunction for the molecular system, and $\{+,-\}$ denotes the space of spin degree of freedom; Ψ belongs to the space of $\mathcal{H}_e \otimes \mathcal{H}_n$, where

$$\mathcal{H}_e = \bigwedge_{i=1}^{N} \mathcal{L}^2(\mathbb{R}^3 \times \{+,-\}, \mathbb{C}),$$

and

$$\mathcal{H}_n = \mathcal{L}^2_{\text{sds}}\big((\mathbb{R}^3 \times A_1) \times \cdots \times (\mathbb{R}^3 \times A_M), \mathbb{C}\big).$$

The $\bigwedge$ symbol denotes the space of antisymmetric functions due to the fermionic property of the electrons, and the "sds" subscript denotes the system-dependent symmetry properties for the nuclei (even number of nuclei: symmetric; odd number of nuclei:antisymmetric). The spin coordinate of the Ith nucleus is denoted by A_I, and the electron spins are denoted by $\{+,-\}$. The square of the magnitude of the wavefunction evaluated at a given spatial and spin coordinates $\{\mathbf{r}_1,\cdots,\mathbf{r}_N;\mathbf{R}_1,\cdots,\mathbf{R}_M;\{+,-\}\}$ represents the probability density of finding the system of nuclei and electrons at $\{\mathbf{r}_1,\cdots,\mathbf{r}_N;\mathbf{R}_1,\cdots,\mathbf{R}_M;\{+,-\}\}$ in $3(M+N)$ spatial dimensions. Hence we require the norm of Ψ in $\mathcal{H}_e\otimes\mathcal{H}_n$ to be 1; in other words, the probability of finding all the nuclei and electron is all of space and any spin coordinates is 1.

$$\|\Psi\|_{\mathcal{H}_e\otimes\mathcal{H}_n}=\sum_{\{+,-\}}\int_{\mathbb{R}^3}\cdots\int_{\mathbb{R}^3}\Psi d\mathbf{r}_1,\cdots d\mathbf{r}_N,d\mathbf{R}_1,\cdots d\mathbf{R}_M=1. \tag{2.2}$$

The operator H in equation (2.1) is the Hamiltonian operator of the molecular system:

$$H=\sum_{i=1}^{N}-\frac{\hbar^2}{2m^e}\Delta_{\mathbf{r}_i}+\sum_{I=1}^{M}-\frac{\hbar^2}{2m_I^n}\Delta_{\mathbf{R}_I}+\sum_{1\le i<j\le N}\frac{q^2}{|\mathbf{r}_i-\mathbf{r}_j|}+\sum_{1\le I<J\le M}\frac{Z_IZ_Jq^2}{|\mathbf{R}_I-\mathbf{R}_J|}-\sum_{i=1}^{N}\sum_{I=1}^{M}\frac{q^2Z_I}{|\mathbf{r}_i-\mathbf{R}_I|}, \tag{2.3}$$

where m^e, m_I^n denote the mass of an electron and the Ith nucleus, respectively; q and Z_Iq denote the charge of the electron and the Ith nucleus; $\mathbf{r}_i$ and $\mathbf{R}_I$ denote the spatial coordinate of the ith electron and the Ith nucleus. The Hamiltonian operator is a self-adjoint operator on the space $\mathcal{H}_e\otimes\mathcal{H}_n$.

We can observe the paralell between the quantum Hamiltonian and the classical Hamiltonian. The following is the kinetic energy operator for the electrons:

$$T_e=\sum_{i=1}^{N}-\frac{\hbar^2}{2m^e}\Delta_{\mathbf{r}_i}, \tag{2.4}$$

the kinetic energy operator for the nuclei:

$$T_n = \sum_{I=1}^{M} -\frac{\hbar^2}{2m_I^n}\Delta_{\mathbf{R}_I},$$

the electrostatic electron-electron repulsion operator:

$$U_{e-e} = \sum_{1\leq i<j\leq N} \frac{q^2}{|\mathbf{r}_i - \mathbf{r}_j|}, \tag{2.5}$$

the electrostatic nucleus-nucleus repulsion operator:

$$U_{n-n} = \sum_{1\leq I<J\leq M} \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|},$$

and the electrostatic nucleus-electron attraction operator:

$$U_{n-e} = -\sum_{i=1}^{N}\sum_{I=1}^{M} \frac{qZ_I}{|\mathbf{r}_i - \mathbf{R}_I|}.$$

The DFT community commonly uses the atomic units where one sets:

$$m^e = 1, q = 1, \hbar = 1.$$

Under this system, the electron-nucleus distance in a Hydrogen atom is of order 1, and its ground-state energy is -0.5. The Hamiltonian operator from (2.3) reduces to

$$H = \sum_{i=1}^{N} -\frac{1}{2}\Delta_{\mathbf{r}_i} + \sum_{I=1}^{M} -\frac{1}{2\alpha_I}\Delta_{\mathbf{R}_I} + \sum_{1\leq i<j\leq N} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} + \sum_{1\leq I<J\leq M} \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} - \sum_{i=1}^{N}\sum_{I=1}^{M} \frac{Z_I}{|\mathbf{r}_i - \mathbf{R}_I|}, \tag{2.6}$$

where $\alpha_I = \frac{m_I^n}{m^e}$.

In practice, we are often interested in finding the ground-state (lowest energy state) wavefunction of the molecular system in equation (2.1), i.e. the smallest eigenvalue and its corresponding eigen-states of the Hamiltonian operator H. The ground-state corresponds

to the wavefunction of the molecular system at 0K. In theory, for a system at non-zero temperature, we should take into account eigen-states of the Hamiltonian with higher energy, known as the excited states. For many applications, the calculation of the ground-state is needed for approximation of the excited states.

Finding the ground-state in equation (2.1) corresponds to finding the infimum of the Rayleigh quotient of H:

$$\epsilon_0 = \inf_{\Psi \in \mathcal{H}_e \otimes \mathcal{H}_n} \langle \Psi | H | \Psi \rangle, \tag{2.7}$$

where $\langle \cdot | \cdot \rangle$ denotes the inner product associated with the space $\mathcal{H}_e \otimes \mathcal{H}_n$.

It would take an audacious scientist to attempt to solve for the eigen-states of the time-independent Schrödinger equation (2.7). The wavefunction Ψ is a function defined on $3(M + N)$ dimension, not counting the spin degree of freedoms. To illustrate the impossibility of solving the Schrödinger equation, suppose that one discretizes each spatial dimension into 100 pieces. The system of equations would involve $100^{3(M+N)}$ degrees of freedom, which equals 10^{600} for a system of 50 nuclei and 50 electrons.

In addition to the large number of dimensions, there is another difficulty associated with the Schrödinger equation written in equation (2.1): according to [12], the Hamiltonian H has a purely continuous spectrum like the quantum position operator X and the quantum momentum operator P. In other words, there is a continuous set of eigen-states. As a result, the infimum in equation (2.7) cannot be attained. To avert this difficulty, physicists came up with an approximation which allows us to separate the nuclei degree of freedom from the electron degree of freedom, and results in a electron Hamiltonian that has a purely discrete spectrum. This approximation is called the Born-Oppenheimer approximation. As a result of the Born-Oppenheimer approximation, we have reduced the quantum degrees of freedom to only those of the electrons.

2.1.1 Born-Oppenheimer Approximation

The key assumption behind the Born-Oppenheimer approximation is that the motion of the nuclei is slow relative to the motion of the electrons, such that at every movement of the nuclei, the electrons have reached their ground-state configuration. In other words, the characteristic time scale to achieve equilibrium for the nucleus is much longer than the characteristic time scale of equilibrium for the electrons. Hence we can treat the spatial coordinates of the nuclei $\{\mathbf{R}_1, \cdots, \mathbf{R}_M\}$ as a parameter, and find the ground-state wavefunction of the electrons for a given set of nuclei coordinates. This assumption is supported by the observation that the mass of a nucleus is at least 1800 times the mass of an electron.

Mathematically, the Born-Oppenheimer approximation allows us to separate the wavefunction Ψ into a *single* product of an electron wavefunction and a nuclei wavefunction:

$$\{\Psi = \Psi_e \Psi_n : \Psi_e \in \mathcal{H}_e, \|\Psi_e\|_{\mathcal{H}_e} = 1, \Psi_n \in \mathcal{H}_n, \|\Psi_n\|_{\mathcal{H}_n} = 1\}.$$

Substitute this approximation into the Rayleigh quotient in equation (2.7), and we get

$$\epsilon_0^{\mathrm{BO}} = \inf_{\Psi_n \in \mathcal{H}_n} \{ \int_{\mathbb{R}^3} \cdots \int_{\mathbb{R}^3} \big(-\frac{1}{2\alpha_I} |\nabla_{\mathbf{R}_I} \Psi_n|^2 + \epsilon_0^e(\mathbf{R}_I, \cdots, \mathbf{R}_M) \big) |\Psi_n|^2) d\mathbf{R}_I \cdots d\mathbf{R}_M \}, \quad (2.8)$$

where

$$\epsilon_0^e(\mathbf{R}_I, \cdots, \mathbf{R}_M) = U_{n-n} + \inf_{\Psi_e \in \mathcal{H}_e} \frac{\langle \Psi_e | H_e | \Psi_e \rangle}{\langle \Psi_e | \Psi_e \rangle}, \quad (2.9)$$

with the electronic Hamiltonian H_e defined by

$$H_e = \sum_{i=1}^{N} -\frac{1}{2} \Delta_{\mathbf{r}_i} + \sum_{1 \le i < j \le N} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} - V_{\mathrm{ext}}(\mathbf{r}_1, \cdots, \mathbf{r}_N, \{\mathbf{R_1}, \cdots, \mathbf{R_M}\}), \quad (2.10)$$

where

$$V_{\text{ext}}(\mathbf{r}_1, \cdots, \mathbf{r}_N, \{\mathbf{R_1}, \cdots, \mathbf{R_M}\}) = \sum_{i=1}^{N} v_{\text{ext}}(\mathbf{r}_i, \{\mathbf{R_1}, \cdots, \mathbf{R_M}\}) = \sum_{i=1}^{N} \Big(\sum_{I=1}^{M} \frac{Z_I}{|\mathbf{r}_i - \mathbf{R}_I|} \Big). \tag{2.11}$$

We will refer to potential due to the nuclei in the electronic problem (2.9) as an external potential; the potential due to the electrons are internal to the problem. The classification of everything that is not electronic potential to be external potential allows us to consider other applied potentials such as electric or magnetic potential on the electronic system in the same generalization. Note that we have adopted a slight abuse of notation where $\langle \cdot | \cdot \rangle$ has been to used to denote both the inner product defined on $\mathcal{H}_e$ and $\mathcal{H}_n$.

In the limit that the mass of the nuclei go to infinity, the kinetic energy of the nuclei can be neglected, and the wavefunction of the nuclei is concentrated on the points $\{\mathbf{R}_I, \cdots \mathbf{R}_M\}$, since the deBroglie wavelength of a nucleus is infinitesimal compared to the deBroglie wavelength of an electron. The infimum problem in equation (2.8) becomes a geometry optimization problem:

$$\epsilon_0^{\text{BO}} = \inf_{\{\mathbf{R}_I, \cdots, \mathbf{R}_M\} \subset \mathbb{R}^{3M}} \epsilon_0^e(\mathbf{R}_I, \cdots, \mathbf{R}_M).$$

The solution of the Schrödinger equation can be solved in two steps: first solve for the electron ground states, by finding the lowest eigenvalue and its corresponding eigenfunction of the electronic Hamiltonian H_e; then solve a geometry optimization problem to get the ground-state energy of the molecular system. The most important consequence of the Born-Oppenheimer approximation is that the electronic Hamiltonian H_e has a purely discrete spectrum, i.e., countable number of eigen-states in many cases. The infimum in the electronic problem in equation (2.9) can be attained depending on the external potential. Although the number of degrees of spatial freedom reduced from $3(M + N)$ to $3N$, the remaining $3N$ dimensions is still impossible to solve directly. This difficulty led to the development of approximate methods like DFT.

2.2 Precursors to density functional theory

The word "density" in density functional theory refers to the electron number density in three dimensions. It is commonly referred to as the electron density of the system; we denote it by $\rho(\mathbf{r})$. One should not confuse the electron density with the probability density in equation (2.2). We will describe subsequently how to obtain the electron density from the probability density given by the electron wavefunction $\Psi_e(\mathbf{r}_1, \cdots, \mathbf{r}_N)$.

In 1964, Kohn and Hohenberg proved that the electronic ground-state energy of the system in equation (2.9) is a unique functional of the electron density derived from ground-state wavefunction. Looking at the minimization problem in equation (2.9); it is not obvious to see why the electron density is relevant. What led Kohn and Hohenberg to the electron density of the system? In fact, Kohn-Hohenberg did not conjure up the concept of the electron density out of nothing. Prior to DFT, there had been a number of approximate methods developed based on relating the electron density to the ground-state energy; these are the Thomas-Fermi models. The motivations for introducing the Thomas-Fermi model in this thesis are two-fold: firstly, it will serve as a transition from the electron ground-state energy as a functional the many-body electron wavefunction $\Psi_e(\mathbf{r}_1, \cdots, \mathbf{r}_N)$ in (2.9) to electron ground state as a functional of the electron density; secondly, the exchange energy functional of the local density approximation, which is a key component of density functional, is taken from the same assumptions of the Thomas-Fermi models. Next we will introduce briefly the Thomas-Fermi models. The spin degree of freedom will be neglected in the following discussion for simplicity.

The Thomas-Fermi model is centered on the problem of *non-interacting* electrons confined in a cubic box of length l; the confinement is imposed through periodic boundary conditions on the many-body electron wavefunction at the boundary of the box. Within the box, the electrons are not subjected to any external potential; in other words, they are "free" electrons in the confined volume. Consider N *non-interacting* electrons confined in the box

as described, and the Hamiltonian of the system inside the box is

$$H\Psi_e(\mathbf{r}_1, \cdots, \mathbf{r}_N) = \sum_{i=1}^{N} \Delta_{\mathbf{r}_i} \Psi_e(\mathbf{r}_1, \cdots, \mathbf{r}_N) = E\Psi_e(\mathbf{r}_1, \cdots, \mathbf{r}_N),$$

subject to the boundary condition,

$$\Psi_e(\mathbf{r}_1, \cdots, \mathbf{r}_N) = \Psi_e(\mathbf{r}_1 + l, \cdots, \mathbf{r}_N + l).$$

Since the Hamiltonian is separable with respect to the spatial coordinate of each electron $\mathbf{r}_i$, the electron wavefunction Ψ_e can be written as a Slater determinant of single electron orbitals [69]:

$$\Psi_e = \frac{1}{\sqrt{N!}} \det \begin{bmatrix} \psi_1(\mathbf{r}_1) & \psi_2(\mathbf{r}_1) & \cdots & \psi_N(\mathbf{r}_1) \\ \psi_1(\mathbf{r}_2) & \psi_2(\mathbf{r}_2) & \cdots & \psi_N(\mathbf{r}_2) \\ \vdots & \vdots & & \vdots \\ \psi_1(\mathbf{r}_N) & \psi_2(\mathbf{r}_N) & \cdots & \psi_N(\mathbf{r}_N), \end{bmatrix} \tag{2.12}$$

where the orbitals $\{\psi_n(\mathbf{r})\}_{n\in\mathbb{Z}}$ are the eigenfunction to the single electron Hamiltonian in a box:

$$H_{\text{box}}\psi_n(\mathbf{r}) = -\frac{1}{2}\Delta_{\mathbf{r}}\psi_n(\mathbf{r}) = \lambda_n\psi_n(\mathbf{r}), \tag{2.13}$$

with the periodic boundary condition:

$$\psi_n(\mathbf{r}) = \psi_n(\mathbf{r} + l).$$

The Slater determinant form ensures that the wavefunction Ψ_e is antisymmetric with respect to exchange of spatial coordinates.

Without considering the boundary conditions, the following solution for the orbitals satisfies the single-electron Hamiltonian in equation (2.13):

$$\psi_{\mathbf{n}}(\mathbf{r}) = C \exp^{i\mathbf{k}\cdot\mathbf{r}},$$

and

$$\lambda_{\mathbf{k}} = \frac{|\mathbf{k}|^2}{2}.$$

As a result of the boundary conditions, the wavefunction $\mathbf{k}$ cannot take arbitrary values; its corresponding wavelength in each spatial dimension has to be an integer multiple of the length of the box. The periodic boundary condition has quantized the wavenumber $\mathbf{k}$. The quantized wave numbers are $\mathbf{k} \equiv [\frac{2\pi n_x}{l}, \frac{2\pi n_y}{l}, \frac{2\pi n_z}{l}]$, x, y, z represent each direction in space, and $\mathbf{n} \equiv [n_x, n_y, n_z]$, and $n_i = \cdots, -2, -1, 0, 1, 2, \cdots$. The corresponding eigenvalues are also quantized according to the quantization of the wave numbers, but since energy levels are proportional to $|\mathbf{k}|^2$, there will be degenerate eigen-states, i.e., wavefunctions that differ in wavenumber but that have the same energy. The electron levels will be filled according to the Pauli-exclusion principle, with only two electrons (assuming a spin-paired system) occupying a given wavefunction with a wavenumber $\mathbf{k}$. The wavefunctions that correspond to the lowest energy will be occupied first at the ground state. The maximum energy reached by a system of N electrons is called the fermi energy, λ_f, and the corresponding magnitude of wavenumber $k_f = |\mathbf{k}_f|$, the fermi wavenumber. We can find what λ_f and k_f for a given system of electrons in a box by arranging all the possible wave numbers in the order of increasing energy, and filling in the states with electrons until we reach N electrons. To find the total energy of the system, which is purely kinetic, we can add up the energy of each electron. In the case where the box is large, i.e., l is very large, and the number of electrons N is also large, we can make an approximation that allows for computation of the fermi level and total energy with far less effort. To illustrate this approximation, let us consider a system of electrons in a box of two dimensions. We can plot the permissible wave numbers as follows: we see in the limit of l is very large, the spacing between consecutive grid points in $\mathbf{k}$-space $\frac{2\pi}{l}$ decreases. The $\mathbf{k}$-space volume occupied by one point (the gray region in Figure 2.1) $\Delta\mathbf{k} = (\frac{2\pi}{l})^2$ also decreases. In this limit, we can approximate the number of grid

points within the circle marked by a radius of k_f in Figure 2.1 by

$$N_p \approx \frac{\pi k_f^2}{\Delta \mathbf{k}}.$$

This approximation improves as the space between grid points decreases. In three dimen-

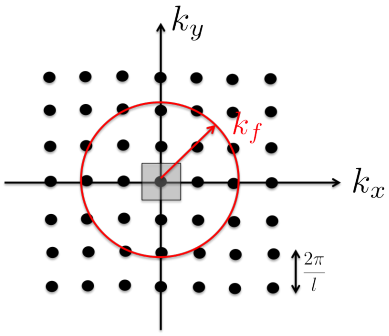

Figure 2.1: **k**-points in two dimensional **k**-space.

sions, we can define fermi wavenumber magnitude by:

$$k_f = \left(\frac{\frac{N}{2}\Delta \mathbf{k}}{\frac{4}{3}\pi}\right)^{1/3} = \left(\frac{3\pi^2 N}{l^3}\right)^{1/3}. \tag{2.14}$$

At this point, we can define a quantity which is going to be the central quantity in DFT, the electron *number* density, or simply the electron density:

$$\rho = \frac{N}{l^3}.$$

We can express both the fermi energy and the fermi wavenumber magnitude as a function of electron density ρ:

$$k_f = (3\pi^2 \rho)^{1/3}. \tag{2.15}$$

and

$$\lambda_f = \frac{(3\pi^2 \rho)^{2/3}}{2}.$$

And to calculate the total (kinetic) energy of the system, the bruit force method would be

$$T = 2\sum_{\mathbf{k}} \lambda(\mathbf{k}) f(\lambda), \tag{2.16}$$

where $f(\lambda)$ is the occupation function of the energy levels:

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

To use the approximation that the box is large, we can write the summation over $\mathbf{k}$ in equation (2.16) as an integral in three dimensions,

$$\begin{aligned} T = 2\sum_{\mathbf{k}} \lambda(\mathbf{k}) f(\lambda) &= \frac{2}{\Delta\mathbf{k}} \sum_{\mathbf{k}} \lambda(\mathbf{k}) f(\lambda) \Delta\mathbf{k} \\ &\approx \frac{l^3}{4\pi^3} \int \lambda(\mathbf{k}) f(\lambda(\mathbf{k})) d\mathbf{k}. \end{aligned} \tag{2.17}$$

Since we know that the energy is only a function of $|\mathbf{k}|$, we can integrate equation (2.17) using spherical coordinates,

$$T(k_f) \approx \frac{l^3}{4\pi^2} \int_0^{k_f} \frac{k^4}{2} \lambda(k) dk.$$

After integration, we can use the relation between k_f and electron density ρ in equation (2.15), and write the total kinetic energy of the system as a function of the electron density:

$$T \approx \frac{3}{10} (3\pi^2)^{2/3} l^3 \rho^{5/3},$$

or kinetic energy *per unit volume*:

$$T_V = \frac{T}{l^3} = \frac{3}{10} (3\pi^2)^{2/3} \rho^{5/3} \equiv C_F \rho^{5/3}. \tag{2.18}$$

The system described above is also called a system of homogeneous electron gas since only homogeneous electron density $\rho = \frac{N}{l^3}$ enters into the equation (2.18). The Thomas-Fermi model approximates the kinetic energy per unit volume of the *inhomogeneous* electron gas by carving up the system into pieces of *locally* homogeneous electron gas, as shown in Figure 2.2. The kinetic energy of the system of the inhomogeneous system is

$$T = \sum_{i=1}^{n} C_F \rho_i^{5/3} V_i.$$

In the limit of the homogenous volumes $V_i \to 0$, the summation becomes an integral; we arrive at the Thomas-Fermi kinetic energy functional, as a function of the electron density:

$$T^{\mathrm{TF}}(\rho) = C_F \int_{\Omega} \rho(\mathbf{r})^{5/3} d\mathbf{r}.$$

It is important to emphasize that the locally homogeneous approximation of the inhomo-

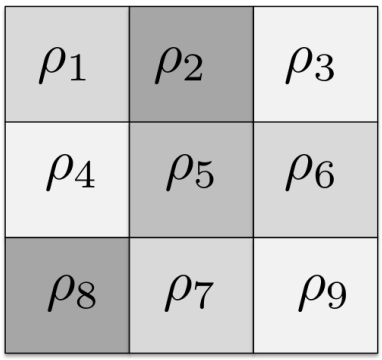

Figure 2.2: A inhomogeneous electron gas divided into pieces of locally homogeneous electron gas.

geneous electron gas is only appropriate when the electron density varies very gradually in space. For instance, this assumption works well for metallic systems where the electrons are not locally bound to any nucleus, but it works poorly for systems with ionic or covalent bonds since the electrons tend to be bound to a given nucleus.

In addition, the Thomas-Fermi model also includes the electron-nuclei, electron-electron

interaction energy,

$$U_{n-e}^{\mathrm{TF}} = \int_{\mathbb{R}^3} \sum_{i=1}^{M} \frac{Z_i}{|\mathbf{R}_i - \mathbf{r}|} \rho(\mathbf{r}) dr,$$

and

$$U_{e-e}^{\mathrm{TF}} = \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r} d\mathbf{r}'.$$

So we have arrived at the Thomas-Fermi energy functional as a function of *only* the electron density $\rho(\mathbf{r})$:

$$E(\rho) = C_F \int_{\mathbb{R}^3} \rho(\mathbf{r})^{5/3} d\mathbf{r} + \int_{\mathbb{R}^3} \sum_{i=1}^{M} \frac{-Z_i}{|\mathbf{R}_i - \mathbf{r}|} \rho(\mathbf{r}) dr + \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r} d\mathbf{r}', \tag{2.19}$$

where $C_F = \frac{3}{10}(3\pi^2)^{2/3}$.

The Thomas-Fermi ground-state energy can be found by minimizing the energy functional in equation (2.19) with the constraint that the total number of electrons is N:

$$\int_{\mathbb{R}^3} \rho(\mathbf{r}) d\mathbf{r} = N.$$

The Thomas-Fermi model remains an academic model because no molecular binding has been predicted by the method [55]. Many improvements and modifications have been made to Thomas-Fermi over the years, but we will not go into detail the different modifications since the focus of this introduction is on DFT. Thomas-Fermi-like models also have been referred to as "orbital"-free DFT since the development of density functional theory. A good introduction to orbital-free DFT is in [19]. With Thomas-Fermi models as a precursor, Kohn and Hohenberg set out to prove rigorously in 1964 the assumption that the ground-state energy of a molecular system can be written *only* as a functional of the electron density.

2.3 Electron density and Hohenberg-Kohn Theorem

Before we state the Hohenberg-Kohn theorem and its proof, we would like to introduce the electron density and its relation to the many-body electron wavefunction Ψ_e.

2.3.1 Electron density

It is quite easy to get an intuitive understanding of what the electron density means physically from the homogenous electron gas; it is less obvious how to find the electron density beyond the homogenous electron gas.

Let us begin by considering $|\Psi_e(\mathbf{x}_1, \cdots, \mathbf{x}_N)|^2$, the *probability density* of finding electron 1 at $\mathbf{x}_1$, electron 2 at $\mathbf{x}_2$, $\cdots$, electron N at $\mathbf{x}_N$, where $\mathbf{x} \equiv (\mathbf{r} : \sigma)$, and $\sigma \in \{+, -\}$. Then

$$\langle \Psi_e | \Psi_e \rangle = \sum_\sigma \underbrace{\int_{R^3} \cdots \int_{\mathbb{R}^3}}_{N} |\Psi_e(\mathbf{r}_1, \cdots, \mathbf{r}_N)|^2 d\mathbf{r}_1, \cdots, d\mathbf{r}_N = 1,$$

is the total probability of finding electron 1 in all of $\mathbb{R}^3$, electron 2 in all of $\mathbb{R}^3$, $\cdots$, electron N in all of $\mathbb{R}^3$. Following suit, we can understanding the following quantity, defined by,

$$\mathfrak{P}_\Omega(\mathbf{r_1}) = \sum_\sigma \int_{\mathbf{\Omega}} \underbrace{\int_{\mathbb{R}^3} \cdots \int_{\mathbb{R}^3}}_{\mathbf{N-1}} |\mathbf{\Psi_e}(\mathbf{r_1}, \mathbf{r_2}, \cdots, \mathbf{r_N})|^\mathbf{2} \mathbf{dr_1}, \cdots, \mathbf{dr_N}, \tag{2.20}$$

as the total probability of finding electron 1 in volume Ω, and electron 2 in all of $\mathbb{R}^3$ $\cdots$, electron N in all of R^3. In other words, independent of the remaining electrons, the probability of finding 1 electron in Ω is $\mathfrak{P}_\Omega$, and I expect to find $\mathfrak{P}_\Omega$ fraction of electron 1 in Ω. Since all the electrons are identical, the total number of electrons we expect to find in Ω is

$$\begin{aligned} n_\Omega &= N\mathfrak{P}_\Omega(\mathbf{r}) \\ &= \int_\Omega \rho(\mathbf{r}) d\mathbf{r}, \end{aligned}$$

where $\rho(\mathbf{r})$ is the electron density of the molecular system, using equation (2.20):

$$\rho(\mathbf{r}) = N \sum_{\sigma} \underbrace{\int_{\mathbb{R}^3} \cdots \int_{\mathbb{R}^3}}_{N-1} |\Psi_e(\mathbf{r}, \mathbf{r}_2, \cdots, \mathbf{r}_N)|^2 d\mathbf{r}_2, \cdots, d\mathbf{r}_N.$$

From the definition of the electron density, we see that the ground-state wavefunction contains more information about the electronic system than the electron density alone. Given a wavefunction, we can always find its corresponding electron density through integration; but given only the electron density, we cannot recover the wavefunction. There may be many wavefunctions that will yield the same electron density.

Next we will state the Hohenberg-Kohn theorem and its proof [33].

2.3.2 Hohenberg-Kohn theorem

Theorem 1 *The ground-state electron density in the electronic problem in equation (2.9) determines uniquely up to a constant the external potential* $V_{\text{ext}}(\mathbf{r}_1, \cdots, \mathbf{r}_N)$ *in equation (2.11) of the system.*

Proof The proof in [33] assumes the non-degeneracy (i.e., uniqueness) of the ground-state wavefunction in equation (2.9), and we will reproduce their proof for completeness. We will discuss later how this assumption can be lifted as result of the work by Lieb, *et al.* [43].

Hohenberg and Kohn proved theorem 1 with proof by contradiction. Suppose for a system of N electrons, there exists two external potentials $V_{\text{ext},1}$, and $V_{\text{ext},2}$ defined by

$$V_{\text{ext},1}(\mathbf{r}_1, \cdots, \mathbf{r}_N) = \sum_{i=1}^{N} v_{\text{ext},1}(\mathbf{r}_i)$$

and

$$V_{\text{ext},2}(\mathbf{r}_1, \cdots, \mathbf{r}_N) = \sum_{i=1}^{N} v_{\text{ext},2}(\mathbf{r}_i).$$

These two potentials differ by more than a constant, and they produce ground-state wave-

functions from equation (2.9) that yield the same electron density.[1] Let us denote the Hamiltonians corresponding to the two external potentials H_1 and H_2, respectively:

$$H_1 = T_e + U_{e-e} + V_{\text{ext},1}$$

and

$$H_2 = T_e + U_{e-e} + V_{\text{ext},2},$$

where T_e and U_{e-e} are defined in equations (2.4) and (2.5). Their corresponding ground-state wavefunctions, $\Psi_{e,1}$ and $\Psi_{e,2}$. Notice that H_1 and H_2 differ only by the external potential. Now consider the variational problem in equation (2.9); for H_1, we have,

$$\begin{aligned} E_1 = \langle \Psi_{e,1}|H_1|\Psi_{e,1}\rangle &= \inf_{\Psi_e \in \mathcal{H}_e, \|\Psi_e\|_{\mathcal{H}_e}=1} \langle \Psi_e|H_1|\Psi_e\rangle \\ < \langle \Psi_{e,2}|H_1|\Psi_{e,2}\rangle &= \langle \Psi_{e,2}|T_e + U_{e-e}|\Psi_{e,2}\rangle + \langle \Psi_{e,2}|V_{\text{ext},1}|\Psi_{e,2}\rangle; \end{aligned}$$

similarly, consider the variation problem (2.9) for H_2:

$$\begin{aligned} E_2 = \langle \Psi_{e,2}|H_2|\Psi_{e,2}\rangle &= \inf_{\Psi_e \in \mathcal{H}_e, \|\Psi_e\|_{\mathcal{H}_e}=1} \langle \Psi_e|H_2|\Psi_e\rangle \\ < \langle \Psi_{e,1}|H_2|\Psi_{e,1}\rangle &= \langle \Psi_{e,1}|T_e + U_{e-e}|\Psi_{e,1}\rangle + \langle \Psi_{e,1}|V_{\text{ext},2}|\Psi_{e,1}\rangle. \end{aligned}$$

Next we take advantage of the fact that H_1 and H_2 only differ by the external potential:

$$\begin{aligned} E_1 &< \langle \Psi_{e,2}|H_1|\Psi_{e,2}\rangle = \langle \Psi_{e,2}|T_e + U_{e_e}|\Psi_{e,2}\rangle + \langle \Psi_{e,2}|V_{\text{ext},1}|\Psi_{e,2}\rangle \\ &= \langle \Psi_{e,2}|T_e + U_{e-e} + V_{\text{ext},2} - V_{\text{ext},2}|\Psi_{e,2}\rangle + \langle \Psi_{e,2}|V_{\text{ext},1}|\Psi_{e,2}\rangle \\ &= E_2 + \langle \Psi_{e,2}|(V_{\text{ext},1} - V_{\text{ext},2})\Psi_{e,2}\rangle, \end{aligned} \tag{2.21}$$

[1] If the potentials differ only by a constant, then the variational problem (2.9) would yield the same ground-state wavefunction, with the ground-state energy differing exactly by the same constant.

and similarly,

$$\begin{aligned} E_2 < \langle \Psi_{e,1} | H_2 | \Psi_{e,1} \rangle &= \langle \Psi_{e,1} | T_e + U_{e_e} | \Psi_{e,1} \rangle + \langle \Psi_{e,1} | V_{\text{ext},2} | \Psi_{e,1} \rangle \\ &= \langle \Psi_{e,1} | T_e + U_{e-e} + V_{\text{ext},1} - V_{\text{ext},1} | \Psi_{e,1} \rangle + \langle \Psi_{e,1} | V_{\text{ext},2} | \Psi_{e,1} \rangle \\ &= E_1 + \langle \Psi_{e,1} | (V_{\text{ext},2} - V_{\text{ext},1}) \Psi_{e,1} \rangle. \end{aligned} \tag{2.22}$$

One can show after some algebra that

$$\langle \Psi_{e,1} | (V_{\text{ext},2} - V_{\text{ext},1}) \Psi_{e,1} \rangle = \int_{\mathbb{R}^3} \big(v_{\text{ext},2}(\mathbf{r}) - v_{\text{ext},1}(\mathbf{r}) \big) \rho(\mathbf{r}) d\mathbf{r} = -\langle \Psi_{e,2} | (V_{\text{ext},1} - V_{\text{ext},2}) \Psi_{e,2} \rangle.$$

Adding equation (2.21) and equation (2.22), we get

$$E_1 + E_2 < E_1 + E_2.$$

Therefore, there cannot exist two external potentials by differing more than a constant that has the same ground-state electron density. ∎

From the Hohenberg-Kohn theorem, given a ground-state electron density, we can determine the number of electrons by integration, and the external potential is determined up to a constant, thus the Hamiltonian is completely determined, and consequently the ground-state energy is completely determined. Further from the variational problem (2.9) for an external potential $V_{\text{ext},1}$,

$$E_1 = \langle \Psi_{e,1} | H_1 | \Psi_{e,1} \rangle = \langle \Psi_{e,1} | T_e + U_{e-e} | \Psi_{e,1} \rangle + \int_{\mathbb{R}^3} v_{\text{ext},1}(\mathbf{r}) \rho(\mathbf{r}) d\mathbf{r},$$

there must exist a functional, $F_{\text{HK}}(\rho)$, such that,

$$F_{\text{HK}}(\rho) = \langle \Psi_{e,1} | T_e + U_{e-e} | \Psi_{e,1} \rangle,$$

where $\Psi_{e,1}$ is the ground-state electron wavefunction for Hamiltonian H_1. The functional F_{HK} is a universal functional,i.e., independent of the external potential of the system; it depends only on the number of electrons in the system N. Hohenberg and Kohn further showed in [33] that there is a variational principle with respect to the electron density for a given external potential:

$$E_0 = \inf_{\rho \in \mathcal{V}} E_{\mathrm{HK}}(\rho) = F_{\mathrm{HK}}(\rho) + \int_{\mathbb{R}^3} v_{\mathrm{ext}}(\mathbf{r})\rho(\mathbf{r})d\mathbf{r}, \tag{2.23}$$

where $\mathcal{V}$ is the space of electron densities that come from ground-state wavefunctions, also known as a V-representable electron densities. There are still two major open questions that remain in the Hohenberg-Kohn energy functional:

1. The exact form of the universal potential $F_{\mathrm{HK}}(\rho)$ is unknown.

2. The necessary and sufficient conditions for the space $\mathcal{V}$ is unknown.

These two open questions render the Hohenberg-Kohn energy functional to a theoretical result; nevertheless, it illuminated a very promising direction for quantum mechanical calculations.

2.4 Kohn-Sham density functional theory

A year later, in 1965, Kohn and Sham [38] came up with an approximation to F_{HK} using the Slater determinant form of electron orbitals in equation (2.12), known as the Kohn-Sham density functional theory. Kohn and Sham sought to solve the first of the two open problems, and neglected the second open problem in their formulation. We restrict the discussion to spin-unpolarized systems for simplicity.

Kohn and Sham approximated $F_{\mathrm{HK}}(\rho)$ by writing down its known contributions, and

leaving the remaining unknown quantities to modeling:

$$F_{\mathrm{HK}}(\rho) = T_0(\rho) + E_{\mathrm{H}}(\rho) + E_{\mathrm{xc}}(\rho). \tag{2.24}$$

The first term in equation (2.24) is the kinetic energy of the electrons if they are non-interacting electrons; the second term,

$$E_{\mathrm{H}}(\rho) = \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|},$$

is the electron-electron repulsion energy if the electrons are classical, also known as the Hartree energy of the system.

The last term contains the remaining interaction energy that has not been accounted for, and it is called the exchange-correlation energy of the system. The exchange-correlation energy functionals were first approximated using the exchange and correlation energies of a locally homogeneous electron gas as described in section 2.2. With these approximations, the Kohn-Sham energy functional becomes

$$E^{\mathrm{KS}}(\rho) = T_0(\rho) + \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} + E_{\mathrm{xc}}(\rho) + \int_{\mathbb{R}^3} v_{\mathrm{ext}}(\mathbf{r})\rho(\mathbf{r})d\mathbf{r}. \tag{2.25}$$

Taking the first variation with respect to the electron density of the Kohn-Sham functional in equation (2.25), subjecting to the constant,

$$\int_{\mathbb{R}^3}\rho(\mathbf{r})d\mathbf{r} = N, \tag{2.26}$$

we arrive at the Euler-Lagrange equation for the Kohn-Sham energy functional:

$$\frac{\delta T_0}{\delta\rho} + \int_{\mathbb{R}^3}\frac{\rho(\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}' + \frac{\delta E_{\mathrm{xc}}}{\delta\rho} + v_{\mathrm{ext}}(\mathbf{r}) + \mu = 0, \tag{2.27}$$

where μ is the Lagrange multiplier for the constraint in equation (2.26).

Kohn and Sham observed that the Euler-Lagrange equation (2.27) for a non-interacting electron system under the external potential is

$$v_{\mathrm{KS}}(\mathbf{r}) = v_{\mathrm{ext}}(\mathbf{r}) + \frac{\delta E_{\mathrm{xc}}}{\delta \rho} + \int_{\mathbb{R}^3} \frac{\rho(\mathbf{r})}{|\mathbf{r} - \mathbf{r}'|}. \tag{2.28}$$

With the assumption that all non-interacting electron systems that are subject to an external potential admit minimizers of the Slate determinant form (2.12), Kohn and Sham came up with an orbital formulation to Hohenberg-Kohn density functional theory. Recall from section 2.2 that the ground-state orbitals of a system of non-interacting electrons can be found by writing the single electron Hamiltonian and selecting its eigenfunctions according to the Pauli-exclusion principle. The corresponding Kohn-Sham single electron Hamiltonian is

$$H^{\mathrm{KS}}\psi_i = \Big(-\frac{1}{2}\Delta + v_{\mathrm{KS}}\big(\rho(\mathbf{r})\big)\Big)\psi_i = \lambda_i^{\mathrm{KS}}\psi_i. \tag{2.29}$$

The corresponding ground-state electron density of the Kohn-Sham system is

$$\rho(\mathbf{r}) = 2\sum_{i=1}^{N/2} |\psi_i(\mathbf{r})|^2, \tag{2.30}$$

where the orbitals ψ_i are the eigenfunctions that correspond to the first $N/2$ lowest eigenvalues. Subsequently, the kinetic energy functional takes the form

$$T_0(\rho) = 2\int_{\mathbb{R}^3} \sum_{i=1}^{N} |\nabla \psi_i(\mathbf{r})|^2 d\mathbf{r}.$$

We can rewrite the Kohn-Sham energy functional in equation (2.25) as a functional of single electron orbitals:

$$E^{\mathrm{KS}}(\rho) = 2\int_{\mathbb{R}^3} \sum_{i=1}^{N/2} |\nabla \psi_i(\mathbf{r})|^2 d\mathbf{r} + \int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r} d\mathbf{r}' + E_{\mathrm{xc}}(\rho) + \int_{\mathbb{R}^3} v_{\mathrm{ext}}(\mathbf{r})\rho(\mathbf{r}), \tag{2.31}$$

subject to the constraint that the orbitals have to be orthonormal:

$$\int_{\mathbb{R}^3} \psi_i(\mathbf{r})\psi_j(\mathbf{r})d\mathbf{r} = \delta_{ij}.$$

Note that the Kohn-Sham single electron Hamiltonian in equation (2.29) is a non-linear functional. The Kohn-Sham potential v_{KS} is a function of the electron density, which is a function of the eigenfunctions of the Hamiltonian in equation (2.30). The solutions of the eigenvalue problem in equation (2.29) can be carried out self-consistently, by starting with an initial guess of electron density ρ_0, and obtaining a Kohn-Sham potential $v_{\mathrm{KS}}(\rho_0)$, and finding the corresponding lowest eigenvalues of $H_{\mathrm{KS}}(\rho_0)$ and then updating the new electron density. This procedure is repeated until a self-consistent density is produced.

2.4.1 Exchange-correlation functional

Since KSDFT is formally exact with the exact exchange-correlation function, the approximation of the exchange-correlation functional is critical to its accuracy. There has been numerous flavors of exchange-correlation functionals developed since 1965. For more information on exchange-correlation functionals, one can refer to [55] and [47]. In their seminal paper [38] Kohn and Sham proposed the local density approximation(LDA). The LDA exchange-correlation functional is based on the inhomogeneous electron gas model as discussed in section 2.2.

The exchange-correlation energy is split into exchange and correlation contributions:

$$E_{\mathrm{xc}}(\rho) = E_{\mathrm{x}}(\rho) + E_{\mathrm{c}}(\rho) = \int_{\mathbb{R}^3} \rho(\mathbf{r})\big(\varepsilon_{\mathrm{x}}(\rho) + \varepsilon_{\mathrm{c}}(\rho)\big)d\mathbf{r}.$$

It is known that the exchange energy is an order of magnitude larger than the correlation energy. In LDA, the exchange energy is computed from plugging in the one-particle density operator $\gamma(\mathbf{r}, \mathbf{r}')$ of the homogenous electron gas, into the exchange energy expression of the Hartree-Fock(HF) approximation. The one-particle density operator is a more general

description of the electron density (see section 2.6.1 for a detailed description), defined by:

$$\gamma(\mathbf{r}, \mathbf{r}^{'}) = 2\sum_{n=1}^{N/2} \psi_n(\mathbf{r})\psi_n^*(\mathbf{r}^{'}).$$

The exchange energy from the HF approximation is

$$E_{\mathrm{x}}(\rho) = \frac{1}{4}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{1}{|\mathbf{r} - \mathbf{r}^{'}|}\gamma(\mathbf{r}, \mathbf{r}^{'})d\mathbf{r}d\mathbf{r}^{'}. \tag{2.32}$$

Recall from section 2.2 the nth orbital for the particle in the box is

$$\psi_n(\mathbf{r}) = \frac{1}{V^{1/2}}\exp(i\mathbf{r}\cdot\mathbf{k}_n).$$

The corresponding one-particle density operator is

$$\gamma(\mathbf{r}, \mathbf{r}^{'}) = \frac{2}{V}\sum_{n=1}^{N} \exp\left(i\mathbf{k}_n\cdot(\mathbf{r} - \mathbf{r}^{'})\right). \tag{2.33}$$

In the limit of the homogenous electron gas (i.e., the limit of $V \to \infty$ and $N \to \infty$ such that $\rho(\mathbf{r}) = \frac{N}{V}$ is finite) we can replace the summation in equation (2.33) with an integral after multiplying by $\frac{\Delta\mathbf{k}}{\Delta\mathbf{k}}$:

$$\gamma(\mathbf{r}, \mathbf{r}^{'}) = \frac{1}{4\pi^3}\int_{\mathbb{R}^3} \exp\left(i\mathbf{k}\cdot(\mathbf{r} - \mathbf{r}^{'})\right). \tag{2.34}$$

Substitute equation (2.34) into the HF exchange energy in equation (2.32), and we simply obtain the exchange energy for the homogeneous electron gas:

$$E_{\mathrm{x}}(\rho) = C_{\mathrm{x}}\int_{\mathbb{R}^3} \rho(\mathbf{r})^{4/3}d\mathbf{r}, \tag{2.35}$$

with $C_{\mathrm{x}} = \frac{3}{4}(\frac{3}{\pi})^{1/3}$. This exchange energy was first calculated by Dirac in [15].

Unlike the exchange energy functional, the correlation energy functional in LDA cannot be obtained exactly in an analytic form. The approximations are often written as a functional

of r_s, defined by:

$$\frac{4}{3}\pi r_s^3 = \frac{1}{\rho}.$$

The different approximations of the correlation functional come from either random phase approximations [81] or numerical calculations of homogenous electron gas in Quantum Monte-Carlo [13]. Since then, more sophisticated exchange-correlation functionals beyond LDA have been introduced in order to increase the accuracy of Kohn-Sham calculations. Some examples include generalized gradient approximation (GGA) [58] and hybrid functionals [8]. We will not go into details on these approximations.

Returning to LDA, the first Kohn-Sham LDA calculations was performed by Tong and Sham in 1966 [76]. Since then, Kohn-Sham density functional theory has become the work-horse of quantum mechanical calculations today.

2.4.2 Pseudopotentials

In the numerical practice of DFT, we often make what is known as the pseudopotential approximation; since it is known that the core electrons in the atom often do not participate in the formation of bonds between atoms, we can assume that the core electron orbitals are "frozen", and can be transferred from a simpler configuration such as a single atom to more complex molecular environments. The adoption of pseudopotential brings two advantages in computation: first, the number of electron orbitals is reduced to only the number of valence electrons in the system; second, the pseudopotential allows us to remove the rapid oscillation of the valence electrons orbitals near the nucleus, which was caused by the orthogonality constraint to the core electron orbitals, hence allowing fewer number of basis to represent the valence electron orbitals in numerical discretization.

The first advantage is evident from the frozen-core approximation, but the second advantage is a result of the observation that scattering can be reproduced over a range of energies by a different potential chosen to have more desirable properties such as smoother orbitals

near the nucleus. The idea of the pseudopotential approximation pre-dates the development of density functional theory, and has been used in many-body wavefunction formulations as well as independent-electron formulations such as Hartree-Fock methods. The concept that led to psedudopotentials used today was the orthogonal plane wave method by Herring in 1940 [30]. The original idea in Herring was to augment the plane-wave basis functions with some other functions that are centered on the nucleus cores so as to reduce the number of plane-wave basis required to represent the valence states. To avoid ill-condition, Herring removed the projection of the nuclei-centered functions from the plane-wave basis:

$$\chi_k^{\mathrm{OPW}}(\mathbf{r}) = \exp(i\mathbf{k}\cdot\mathbf{r}) - \sum_j^m \langle w_j|\mathbf{k}\rangle w_j(\mathbf{r}),$$

where

$$\langle w_j|\mathbf{k}\rangle = \int_{\mathbb{R}^3} w_j(\mathbf{r}) \exp(i\mathbf{k}\cdot\mathbf{r}) d\mathbf{r}.$$

The choice of the nuclei-centered functions is critical to the success of the OPW method. Herring chose a function that obeys wavefunctions of the form

$$-\frac{1}{2}\Delta_{\mathbf{r}} w_j(\mathbf{r}) + V_j(\mathbf{r}) w_j(\mathbf{r}) = E_j w_j(\mathbf{r}).$$

In short, the OPW formulation is nothing but writing the valence orbitals as a linear combination of a smoothed function and a few nuclei-centered functions:

$$\psi^v(\mathbf{r}) = \tilde{\psi}^v(\mathbf{r}) + \sum_j^m c_j w_j(\mathbf{r}). \tag{2.36}$$

In 1959, Phillips and Kleinnman [61] adopted the OPW formulation to independent-orbital approximations, and derived formally a pseudopotential approximation that contains a non-local potential. They substituted equation (2.36) into the single-particle Schrödinger equation, with $w_j(\mathbf{r}) = \psi_j^c(\mathbf{r})$, where $\psi_j^c(\mathbf{r})$ are the core electron orbitals of the reference

system used to create the pseudopotential:

$$H\psi_j^c = \lambda_j^c \psi_j^c,$$

and

$$H\psi^v = \lambda^v \psi^v.$$

Using the fact that the valence orbitals are orthogonal to the core orbitals, in bra-ket notation:

$$\begin{aligned} \langle \psi_i^c | \psi^v \rangle &= \langle \psi_i^c | \tilde{\psi}^v \rangle + \sum_j^m c_j \langle \psi_i^c | \psi_j^c \rangle = 0 \\ \implies c_i &= -\langle \psi_i^c | \tilde{\psi}^v \rangle, \end{aligned}$$

we can derive a Hamiltonian that yields $\tilde{\psi}^v$ as an eigenfunction:

$$\begin{aligned} H|\psi^v\rangle &= H|\tilde{\psi}^v\rangle + \sum_{j=1}^m c_j H|\psi_j^c\rangle = \lambda^v \psi^v\rangle \\ &= H|\tilde{\psi}^v\rangle - \sum_{j=1}^m \lambda_j^c |\psi_j^c\rangle\langle\psi_j^c|\tilde{\psi}^v\rangle \\ &= \lambda^v \underbrace{\left(|\tilde{\psi}^v\rangle + \sum_{j=1}^m |\psi_j^c\rangle\langle\psi_j^c|\tilde{\psi}^v\right)\rangle}_{\psi^v} \\ \implies H^{\text{ps}}\tilde{\psi}^v &= \lambda^v \tilde{\psi}^v, \end{aligned}$$

where

$$H^{\text{ps}} = H + \sum_j^m (\lambda^v - \lambda_j^c) |\psi_j^c\rangle\langle\psi_j^c|. \tag{2.37}$$

The potential in equation (2.37) is a repulsive potential because $\lambda^v - \lambda_j^c$ is a positive quantity for all j, hence giving us a weaker attractive potential than the original potential. The resulted pseudopotential is nonlocal, which means that it *cannot* be written in the form

$V^{\text{ps}}(\mathbf{r}, \mathbf{r}') = v^{\text{ps}}(\mathbf{r})\delta(|\mathbf{r} - \mathbf{r}'|)$.

In practice, this is *not* how pseudopotential is constructed, but they retain the same non-local structure. It's worthwhile to point out that $\psi^v, \{\psi_j^c\}_{j=1}^m$, and $\tilde{\psi}^v$ are eigenfunctions of H^{ps} with the corresponding eigenvalue λ^v. In addition, this pseudopotential contains the core orbitals, which are high oscillatory; it also contains the original potential V in $H = -\frac{1}{2}\Delta + V$, which has a singularity at the location of the nucleus.

Many of the pseudopotentials that are used in practice are so called the "norm-conserving" pseudopotentials, which has to satisfy the following four conditions:

1. All-electron and pseudo-valence eigenvalues agree for the chosen atomic reference configuration.

2. All-electron and pseudo-valence wavefunctions agree beyond a chosen core radius R_c.

3. The logarithmic derivatives of the all-electron and pseudo-wavefunctions agree at R_c.

4. The first energy derivative of the logarithmic derivatives of the all-electron and pseudo-wavefunctions agree at R_c, and therefore for all $r \geq R_c$.

These four conditions were given by Hamann, Schluter, and Chiang in 1979 [25]. The last three conditions ensure a good transferability of the pseudopotential, as well as allowing flexibility to smooth the core region of the valence orbitals. A common pseudopotential used in practice was developed by Troullier and Martin in [78]. Another type of pseudopotential that is common use are the ultra-soft pseudopotentials that relax the norm-conservation constraint, developed by Vanderbilt in 1990 [79].

2.5 Density functional theory made more rigorous by Levy and Lieb

In 1965, Kohn-Sham left open several mathematical questions. The question regarding the space of ground-state electron densities raised in section 2.3 was solved by Levy and Lieb in

1982 ([41] and [43]).

The names Levy and Lieb are mentioned far less frequently than their contributions would merit in the DFT community. Levy and Lieb build a firm mathematical foundation for DFT that justifies the Kohn-Sham approximations. They made two key contributions in 1982:

1. They removed the restriction on the non-degeneracy of the ground-state wavefunction assumed by Hohenberg and Kohn.

2. They removed the constraint that the space of the electron densities has to be *ground-state* electron densities. They rigorously proved the existence of an energy functional that is defined over a space of densities for which we know the necessary and sufficient conditions.

We will now explain the contributions of Levy and Lieb in more detail, starting from the electronic variational problem in equation (2.9). To find the ground-state energy of the electronic problem, we have to search over the entire space of antisymmetric N-electron wavefunctions in the space $\mathcal{H}_e$. Levy proposed to break $\mathcal{H}_e$ into groups of antisymmetric wavefunctions that have the same density, look for the minimum of equation (2.9) within a given electron density group, and then minimize over all the possible electron densities. A good analogy of this search method is given by [55]: suppose we are interested in finding the tallest student in a high school. Instead of making every student in the school line up in the order of heights, we can ask each class to find the tallest student in their class, and then lastly look for the tallest student out of the tallest student from each class.

Going back to the electronic problem, mathematically, we have

$$\begin{aligned}\epsilon_0 &= \inf_{\Psi_e \in \mathcal{H}_e} \langle \Psi_e | H | \Psi_e \rangle \\ &= \inf_{\rho \in \mathcal{N}} \{ \inf_{\Psi_e \to \rho} \langle \Psi_e | T_e + U_{e-e} | \Psi_e \rangle + \int_{\mathbb{R}^3} v_{\text{ext}}(\mathbf{r}) \rho(\mathbf{r}) d\mathbf{r} \}, \end{aligned} \tag{2.38}$$

where $\Psi_e \to \rho$ means all the antisymmetric wavefunctions in $\mathcal{H}_e$ that yield electron density ρ, and $\mathcal{N}$ denotes the space of electron densities that come from antisymmetric wavefunction of an N-particle system, which contains the space of *ground-state* electron densities $\mathcal{V}$ in the Hohenberg-Kohn theorem. Most importantly, the necessary and sufficient conditions for the space $\mathcal{N}$ is known. The conditions are

$$\int_{\mathbb{R}^3} \rho(\mathbf{r})d\mathbf{r} = N, \rho(\mathbf{r}) \geq 0, \int_{\mathbb{R}^3} |\nabla\sqrt{\rho(\mathbf{r})}|^2 d\mathbf{r} < \infty.$$

This space is known as the N-representable densities.

Hence we can define the Levy-Lieb universal functional $F_{LL}(\rho)$:

$$F_{LL}(\rho) = \inf_{\Psi_e \to \rho} \langle \Psi_e | T_e + U_{e-e} | \Psi_e \rangle. \tag{2.39}$$

Lieb shows the existence of minimizers for $F_{LL}(\rho)$ in [43]. We can split $F_{LL}(\rho)$ into the kinetic energy functional and the coulomb-interaction functional by writing

$$F_{LL}(\rho) = T_e(\rho) + U_{e-e}(\rho),$$

where

$$T_e(\rho) = \langle \Psi^\rho_{e,\min} | T_e | \Psi^\rho_{e,\min} \rangle,$$

and

$$U_{e-e}(\rho) = \langle \Psi^\rho_{e,\min} | U_{e-e} | \Psi^\rho_{e,\min} \rangle,$$

with $\Psi^\rho_{e,\min}$ being a minimizer to equation (2.39).

Putting everything together, we have the Levy-Lieb energy functional:

$$\epsilon_0 = \inf_{\rho \in \mathcal{N}} \{F_{LL}(\rho) + \int_{\mathbb{R}^3} v_{\text{ext}}(\mathbf{r})\rho(\mathbf{r})d\mathbf{r}\}.$$

With a more rigorous definition of the universal electronic functional $F_{LL}(\rho)$, we can

follow suit in the Kohn-Sham formulation to construct the Levy-Lieb universal functional for a system of N non-interacting electrons. The Hamiltonian H_0 for the N non-interacting electrons is

$$H_0 = -\frac{1}{2}\sum_{i=1}^{N}\Delta_{\mathbf{r}_i} + \sum_{i=1}^{N} v_{\text{ext}}(\mathbf{r}_i).$$

Consequently, the universal functional for the independent electron system consists of only the kinetic energy

$$T_0(\rho) = \inf_{\Psi_e \to \rho} \langle \Psi_e | -\frac{1}{2}\sum_{i=1}^{N}\Delta_{\mathbf{r}_i} | \Psi_e \rangle. \tag{2.40}$$

When the variational problem in equation (2.40) admits a minimizer in the form of a Slater determinant as shown in equation (2.12), the kinetic energy functional simplifies to

$$T_0(\rho) = \inf_{\mathcal{I}_N} \frac{1}{2}\sum_{i=1}^{N} |\nabla\psi_i|^2, \tag{2.41}$$

where $\mathcal{I}_N = \{\psi \in H^1(\mathbb{R}^3), \int_{\mathbb{R}^3} \psi_i\psi_j = \delta_{ij}$, and $\sum_{i=1}^{N} |\psi_i(\mathbf{r})|^2 = \rho(\mathbf{r}).\}$. However, not all ground-state non-interacting electron density admits a Slater determinant minimizer, so the orbital formulation of Kohn-Sham density functional theory in section 2.4 constrains the search space to only Slater-determinant representable electron densities, and hence is a strict upper bound to the exact ground state energy.

2.6 Extended Kohn-Sham Energy Functional

To avoid the representation difficulty in the orbital formulation of the Kohn-Sham energy functional, Lieb [43] proposed a density functional that has a precise mathematical description. The Lieb density functional F_L was formulated using N-particle density operator, Γ_N, which is a linear operator on $\mathcal{H}_e$. We will introduce the density operator before we derive the Lieb density functional.

2.6.1 Density Operator

In quantum mechanics, the density operator is a more general description of the electronic system. Whenever the state of an electronic system can be described by a wavefunction, then the system is in a *pure* state. When an electronic system cannot be described by any wavefunction, (e.g., when the system is a sub-system of a larger system, and it doesn't have a Hamiltonian containing only its own degree of freedom), then the system is in a *mixed* state. A system in a mixed state has to be described using a density operator, whereas a system in a pure state can be represented using either a wavefunction or density operator.

Suppose a system of N-electrons are in the state $\Psi_e(\mathbf{r}_1, \cdots, \mathbf{r}_N)$; then the N-particle density operator that describes the system is

$$\Gamma_N = |\Psi_e\rangle\langle\Psi_e|.$$

Notice that even though the wavefunction $\Psi_e(\mathbf{r}_1, \cdots, \mathbf{r}_N)$ is only unique up to a phase shift, the N-particle density operator is completely unique for a given electronic system. In the pure state, the N-particle density operator is idempotent, i.e., $\Gamma_N^2 = I$ due to the normalization of the wavefunction Ψ_e. An the expectation value of a given operator A on the pure-state electron system can be written as

$$\langle A\rangle = \mathrm{Tr}(A\Gamma_N) = \langle\Psi_e|A|\Psi_e\rangle.$$

A system of N particles in a mixed state can be written as a sum of the probabilities of finding the particles in a given pure-state, $\Psi_{e,i}$:

$$\Gamma_{(N,\mathrm{mixed})} = \sum_{i=1}^{\infty} p_i|\Psi_{e,i}\rangle\langle\Psi_{e,i}|,$$

where $\{\Psi_{e,i}\}$ is orthonormal, and the p_i are probabilities:

$$p_i \geq 0, \sum_{i=1}^{\infty} p_i = 1.$$

It is evident that the pure-state N-particle density operator is a special case of the mixed-state density operator with one of the $p_j = 1$, and the remainder $p_{i\neq j} = 0$.

2.6.2 Extended Kohn-Sham Energy Functional

Using the N-particle density operator, the ground-state energy in equation (2.7) is equivalent to

$$\epsilon_0 = \inf_{\Gamma_{(N,\text{mixed})} \in \mathcal{D}_N} \text{Tr}(H\Gamma_{(N,\text{mixed})}) = \sum_{i=1}^{\infty} p_i \langle \Psi_{e,i}|H|\Psi_{e,i}\rangle,$$

where $\mathcal{D}_N = \{\Gamma = \sum_{i=1}^{\infty} p_i |\Psi_{e,i}\rangle\langle\Psi_{e,i}|, 0 \leq p_i \leq 1, \sum_{i=1}^{\infty} p_i = 1, \Psi_{e,i} \in \mathcal{H}_e\}$ is the set of mixed-state N-particle density operators.

We can define an analogous universal functional to the Levy-Lieb universal functional using the mixed-state density operators in $\mathcal{D}_N$. This is known as the Lieb functional:

$$F_L(\rho) = \inf_{\Gamma_{(N,\text{mixed})} \to \rho} \text{Tr}((T_e + U_{e-e})\Gamma_{N,\text{mixed}}),$$

where $\Gamma_{(N,\text{mixed})} \to \rho$ are the mixed-state density operators $\Gamma_{(N,\text{mixed})} \in \mathcal{D}_N$ that have electron density ρ. When we write the Lieb universal functional for a system of non-interacting electrons, we can define the Janak kinetic energy functional as

$$T_J(\rho) = \inf_{\Gamma_{(N,\text{mixed})} \to \rho} \text{Tr}(H_0\Gamma_{(N,\text{mixed})}) = \inf_{\Gamma_{(N,\text{mixed})} \to \rho} \{-\frac{1}{2}\text{Tr}(\sum_{i=1}^{N} \Delta_{\mathbf{r}_i}\Gamma_{(N,\text{mixed})})\}. \tag{2.42}$$

With some algebra, we can show that for any mixed-state N-particle density operator,

$$\text{Tr}(H_0\Gamma_{(N,\text{mixed})}) = -\frac{1}{2}\text{Tr}(\Delta\gamma),$$

where γ is the *one-particle reduced* density operator associated with $\Gamma_{N,\text{mixed}}$ defined by

$$\gamma(\mathbf{r}_1, \mathbf{r}_1^{'}) = N \underbrace{\int_{\mathbb{R}^3} \cdots \int_{\mathbb{R}^3}}_{N-1} \sum_{i=1}^{\infty} p_i \Psi_{e,i}^{*}(\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_N) \Psi_{e,i}(\mathbf{r}_1^{'}, \mathbf{r}_2^{'}, \cdots, \mathbf{r}_N^{'}) d\mathbf{r}_2 d\mathbf{r}_2^{'} \cdots d\mathbf{r}_N d\mathbf{r}_N^{'}.$$

Further, we know a lot about the space of the one-particle reduced density operator that derives from the space of mixed-state N-particle density operators: the one-particle reduced density operators are completely described by

$$\mathcal{X}_N = \{\gamma = \sum_{i=1}^{\infty} n_i \psi_i(\mathbf{r}) \psi_i(\mathbf{r}^{'}), \psi_i \in H^1(\mathbb{R}^3), \int_{\mathbb{R}^3} \psi_i \psi_j d\mathbf{r} = \delta_{ij}, 0 \leq n_i \leq 1, \sum_{i=1}^{\infty} n_i = N\}.$$

We can define an electron density from every $\gamma \in \mathcal{X}_N$:

$$\rho(\mathbf{r}) = \gamma(\mathbf{r}, \mathbf{r}) = \sum_{i=1}^{\infty} |\psi_i(\mathbf{r})|^2. \tag{2.43}$$

Using this description of $\mathcal{X}_N$, the Janak kinetic energy functional in equation (2.42) simplifies to

$$T_J(\rho) = \inf_{\gamma \in \mathcal{X}_N} \frac{1}{2} \sum_{i=1}^{\infty} n_i \int_{\mathbb{R}^3} |\nabla \psi_i|^2 d\mathbf{r}.$$

Following Kohn-Sham's definition of the exchange-correlation functional, we have

$$E_{\text{xc}}(\rho) = F_L(\rho) - T_J(\rho) - E_{\text{H}}(\rho).$$

We have now derived the *extended* Kohn-Sham model:

$$E_0^{EKS} = \inf_{\gamma \in \mathcal{X}_N} \{-\frac{1}{2}\text{Tr}(\Delta\gamma) + \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(\mathbf{r})\rho(\mathbf{r}^{'})}{|\mathbf{r} - \mathbf{r}^{'}|} d\mathbf{r} d\mathbf{r}^{'} + \int_{\mathbb{R}^3} \rho(\mathbf{r}) v_{\text{ext}}(\mathbf{r}) d\mathbf{r} + E_{\text{xc}}(\rho)\}. \tag{2.44}$$

Unlike the Kohn-Sham model, the extended Kohn-Sham model is defined over a space of well-defined solutions $\mathcal{X}_N$, and it enables validation as well as verification of the model.

Chapter 3

Linear-scaling methods in density functional theory

As we introduced in chapter 2, for a given set of nuclei positions $\{\mathbf{R_1}, \cdots, \mathbf{R_M}\}$, finding the ground-state energy of the Kohn-Sham energy functional consists of solving the non-linear eigenvalue problem:

$$H^{\mathrm{KS}}\psi_i(\mathbf{r}) = \{-\frac{1}{2}\Delta + v_{\mathrm{KS}}(\rho)\}\psi_i(\mathbf{r}) = \lambda_i\psi_i(\mathbf{r}), \tag{3.1}$$

where the non-linearity lies in the effective Kohn-Sham potential defined by

$$v_{\mathrm{KS}}(\rho) = \int_{\mathbb{R}^3} \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}' + v_{\mathrm{ext}}(\mathbf{r}) + v_{\mathrm{xc}}(\rho)$$

and

$$v_{\mathrm{xc}}(\rho) = \frac{\partial E_{\mathrm{xc}}(\rho)}{\partial \rho}.$$

The Kohn-Sham ground-state energy for a spin-unpolarized molecular system equals

$$\begin{aligned}
&\epsilon_0^{\mathrm{KS}}(\{\mathbf{R_1}, \cdots, \mathbf{R_M}\}) \\
&= \sum_{i=1}^{N/2} |\nabla\psi_i(\mathbf{r})|^2 + \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}d\mathbf{r}' + \int_{\mathbb{R}^3} v_{\mathrm{ext}}(\mathbf{r}, \{\mathbf{R_1}, \cdots, \mathbf{R_M}\})\rho(\mathbf{r})d\mathbf{r} + E_{\mathrm{xc}}(\rho) + U_{n-n}(\{\mathbf{R_1}, \cdots, \mathbf{R_M}\}) \\
&= 2\sum_{i=1}^{N/2} \lambda_i - \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}d\mathbf{r}' - \int_{\mathbb{R}^3} v_{\mathrm{xc}}(\mathbf{r})\rho(\mathbf{r})d\mathbf{r} + E_{\mathrm{xc}}(\rho) + U_{n-n}(\{\mathbf{R_1}, \cdots, \mathbf{R_M}\}),
\end{aligned}$$

where $\{\lambda_i\}_{i=1}^{N/2}$ and $\{\psi_i(\mathbf{r})\}_{i=1}^{N/2}$ are the lowest $N/2$ eigenvalues and the corresponding eigenfunctions of in equation (3.1), and the electron density $\rho(\mathbf{r})$ is defined by

$$\rho(\mathbf{r}) = 2\sum_{i=1}^{N/2} |\psi(\mathbf{r})|^2. \tag{3.2}$$

The eigenvalues correspond to the energy of each Kohn-Sham electron; it is important to emphasize that the Kohn-Sham electrons are not exactly like the electrons that are in the molecular system. The Kohn-Sham electrons do not interact with one another; they interact only with the effective potential.

The conventional solution to the Kohn-Sham equations is the direct diagonalization of the discretized Kohn-Sham Hamiltonian matrix; the computational cost of diagonalization scales cubically with respect to the number of electrons in the system:

$$C = a_1 N^3.$$

The cubic-scaling cost of the diagonalizing procedure has been a bottle-neck to applying KSDFT to molecular systems larger than a thousands of atoms. When the system size doubles, the computation cost jumps 8-fold. This difficulty led to the development of linear-scaling implementations of KSDFT, with computational cost that increases linearly with respect to the system size:

$$C = a_2 N.$$

The linear-scaling methods avoid the diagonalization of the discretized Kohn-Sham Hamiltonian matrix; they either take advantage of the localization properties of the electron orbitals in certain types of materials and/or the sparsity of the Hamiltonian matrix to approximate the ground-state energy of the Kohn-Sham system. The prefactor a_2 in linear scaling methods are always larger than the prefactor a_1 in diagonalization methods, thereby causing a cross-over point in the number of atoms, beyond which the linear-scaling methods will be

cheaper computationally.

There are many flavors of linear-scaling implementations, but they are broadly divided into two main categories: the first category approximates the density matrix, the finite dimension realization of the one-particle density operator as defined in section 2.6.1; these methods are known as density matrix expansion methods in literature. The second category approximates the occupied orbitals iteratively. We will describe in details several examples of density matrix expansion methods and briefly describe the methods from the second category. There are several excellent reviews on linear-scaling methods in DFT ([23],[10], and [47]).

3.0.3 Density matrix expansion methods

The basis for density matrix expansion methods lies in the observation that the ground-state one-particle density operator γ shares the same complete set of eigen-states with the Kohn-Sham Hamiltonian operator H^{KS}. This observation can be seen in the definition of the ground-state electron density in equation (3.2) and in equation (2.43); the ground-state Kohn-Sham one-particle density operator has eigenvalue 1 for the occupied eigen-states, and eigenvalue 0 for the unoccupied eigen-states.

$$\gamma(\mathbf{r}, \mathbf{r}^{'}) = 2 \sum_{i=1}^{N/2} \psi_i(\mathbf{r}) \psi_i(\mathbf{r}^{'}). \tag{3.3}$$

The electron density defined by the density operator in equation (3.3) is

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

Using spectral theorem from the theory of self-adjoint operators [65], we can write the density operator as a function of the Hamiltonian operator.

Theorem 2 *Let P be the resolution of the identity associated with H^{KS}, then for every*

bounded Borel function on $\sigma(H^{\mathrm{KS}})$, the spectrum of the Kohn-Sham Hamiltonian, it can be written in the form,

$$f(H^{\mathrm{KS}}) = \int_{\sigma(H^{\mathrm{KS}})} f(\lambda) dP(\lambda).$$

In other words, we can write the one-particle density operator as

$$\gamma = g(H^{\mathrm{KS}}).$$

The finite dimensional realization of spectral theorem is the spectral decomposition of hermitian matrices in linear algebra. For every hermitian matrix H, we can define its spectral decomposition:

$$H = \sum_{i=1}^{N_d} \lambda_i \psi_i \otimes \psi_i,$$

where λ_i and ψ_i are the eigenvalues and eigenvectors of the matrix H, and N_d is the size of the matrix H. We can define a matrix function $g(H)$ as

$$g(H) = \sum_{i=1}^{N_d} g(\lambda_i) \psi_i \otimes \psi_i.$$

The density matrix, can defined as the matrix function $g(H^{\mathrm{KS}})$:

$$g(\lambda) = 2 \begin{cases} 1, & \text{if } \lambda \leq \lambda_{N/2}, \\ 0, & \text{otherwise}, \end{cases} \tag{3.4}$$

where $\lambda_{N/2}$ is the $N/2$ eigenvalue of the Hamiltonian matrix.

In the extended Kohn-Sham energy functional in equation (2.44), the occupation number of the Kohn-Sham orbitals can take fractional occupations, and the density matrix can be defined as

$$g(\lambda) = 2 \begin{cases} 1, & \text{if } \lambda \leq \lambda_f, \\ 0, & \text{otherwise}, \end{cases} \tag{3.5}$$

where λ_f is the energy of the system. It is defined so that the total number of electrons in

the molecular system is conserved:

$$\mathrm{Tr}(\gamma) = \mathrm{Tr}\big(g(H^{\mathrm{KS}})\big) = \int_{\mathbb{R}^3} \rho(\mathbf{r}) d\mathbf{r} = N.$$

It is important to emphasize here that to evaluate the density matrix exactly would involve finding a spectral decomposition of the Hamiltonian matrix, which would incur cubic scaling computational cost. The key intuition behind density matrix expansion methods is that we can approximate the ground-state density matrix by using simpler functions of the Kohn-Sham Hamiltonian matrix that can be computed at linear cost:

$$g(H^{\mathrm{KS}}) \approx \sum_{i=1}^{n_p} c_i p_i(H^{\mathrm{KS}}).$$

We will refer to these simpler functions as basis functions on the spectrum. There are several variations of the spectral basis functions, e.g., polynomial functions and rational functions. I will describe a few examples of the density matrix expansion methods and their algorithms.

3.0.3.1 Chebyshev polynomials

Polynomial approximations of the density matrix was first introduced by Goedecker and Colombo in [20]; since then, there have been numerous adaptations of polynomial approximations (e.g., [22], [6], and [73]). We will introduce in detail here the polynomial approximation using Chebyshev polynomials.

Chebyshev polynomials $\{T_j\}_{j=1}^{\infty}$ are orthogonal polynomials with respect to the weight function [24],

$$w(x) = (1-x)^{-\frac{1}{2}}(1+x)^{-\frac{1}{2}}.$$

They satisfy the following 3-term recursion relation,

$$T_0(x) = 1,$$

$$T_1(x) = x,$$

$$T_{j+1}(x) = 2xT_j(x) - T_{j-1}(x).$$

They form a complete basis for the inner product space $\mathcal{L}^2\big([-1,1], w(x)\big)$, and every function $f(x) \in \mathcal{L}^2\big([-1,1], w(x)\big)$ can be written as

$$f(x) = \sum_{j=1}^{\infty} c_j T_j(x).$$

The coefficients of expansion can be found by taking the inner product,

$$c_j = \int_{\mathbb{R}^3} f(x) T_j(x) w(x) dx.$$

In application to DFT, the key idea is that we can approximate the density matrix using a truncated Chebyshev polynomial expansion in the Hamiltonian matrix:

$$\gamma = \sum_{j=1}^{n_p} c_j T_j(H).$$

Since the Chebyshev polynomials are only dense for functions with domain $[-1.1]$, we first have to transform the discretized Hamiltonian matrix so that its spectrum falls completely within $[-1.1]$. This transformation requires an estimate of the largest and the smallest eigenvalue of the Hamiltonian matrix. The methods used by Goedecker *et al.* is to use a Chebyshev filter where one constructs a Chebyshev polynomial fit $p_{\text{up}}(\lambda)$ of a function that vanishes below some λ_{max}, but blows up for energies larger than λ_{max}. If $\text{Tr}(p_{\text{up}}(H))$ does not vanish then we have non vanishing eigenvalues beyond λ_{max}. The same procedure can be used to find λ_{min}.

The density matrix written as a matrix function of the Hamiltonian matrix is the step function defined in equation (3.5). Due to the discontinuity of the matrix function at λ_N, Chebyshev approximation suffers from Gibbs oscillations near the discontinuity [23]; there-

fore, in numerical practice one has to regularize the discontinuity in equation (3.5). In [20], the authors took the Chebyshev expansion of the Fermi-Dirac distribution:

$$g_{\text{fermi}}(\lambda) = \frac{1}{1 - \exp(\frac{\lambda - \lambda_f}{k_{\text{B}}T})}, \tag{3.6}$$

where k_{B} is the Boltzmann constant, T is the electronic temperature, and λ_f is the fermi energy, defined by the number of electrons in the molecular system:

$$\text{Tr}(g_{\text{fermi}}(H^{\text{KS}})) = N.$$

The Fermi-Dirac distribution describes the distribution of N identical particles subject to the Pauli-exclusion principle in thermo-equilibrium. This is the reason why density matrix expansion methods are also called Fermi-operator expansion methods in literature. Of course the choice of the regularization is by no means is unique, and the same authors also suggested using the erf function:

$$f(\lambda) = \frac{1}{2}\{1 - \text{erf}\big((\lambda - \lambda_f)/\Delta\lambda\big)\},$$

where $\Delta\lambda$ is chosen for numerical convenience, and it serves the same role as $k_{\text{B}}T$ in the Fermi-Dirac distribution.

When Goedecker and Colombo first developed the Chebyshev expansion of the density matrix, they believed that the way to achieve linear-scaling computation cost was through taking advantage of the decay properties of the density operator in the spatial $\mathbf{r}$-basis, in addition to taking advantage of the sparsity of the Hamiltonian matrix. We will see later that the decay property of the density operator is not necessary.

The density operator in spatial coordinates decay algebraically in metals at zero temperature [46]:

$$\gamma(\mathbf{r}, \mathbf{r}^{'}) \propto k_f \frac{cos(k_f|\mathbf{r} - \mathbf{r}^{'}|)}{|\mathbf{r} - \mathbf{r}^{'}|^2},$$

where k_f is the fermi wave number magnitude described in equation (2.14). With the intro-

duction of a finite temperature, T, equivalent to regularization of the density matrix using the Fermi-Dirac distribution in equation (3.6), the density operator in spatial coordinates exhibits exponential decay ([21] and [34]):

$$\gamma(\mathbf{r}, \mathbf{r}^{'}) \propto k_f \frac{cos(k_f|\mathbf{r} - \mathbf{r}^{'}|)}{|\mathbf{r} - \mathbf{r}^{'}|^2} \exp\big(- c\frac{k_{\mathrm{B}}T}{k_f}|\mathbf{r} - \mathbf{r}^{'}|\big),$$

where c is a constant on the order of 1. In insulators, we can adapt the same finite-temperature density operator as long as $k_{\mathrm{B}}T$ is less than the band gap of the material.

In order for the finite dimensional approximation of the density operator, the density matrix, to reflect the decay properties of the density operator $\gamma(\mathbf{r}, \mathbf{r}^{'})$, one has to use a localized basis to discretize the Hamiltonian matrix and the density matrix (e.g., finite element methods, atom-centered Gaussian-type basis, finite difference method, etc). Other than finite-difference methods, the other localized bases mentioned above are not orthonormal. In the following discussion, we will assume that the localized basis are orthonormal. Given that we use N_d number of basis functions, the density matrix is a $N_d \times N_d$ matrix. Using the Chebyshev polynomial approximation, the computation cost for each column of the density matrix can be evaluated using a recursive relation; each column of the polynomial matrix function also obeys the recursive relation. Let t_l^j denote the lth column of the Chebyshev matrix $T_j(H)$ and e_l denote the unit vector with 1 at the lth entry and zeros in all remaining entries.

$$|t_l^0\rangle = |e_l\rangle,$$

$$t_l^1\rangle = H|e_l\rangle,$$

$$t_l^{j+1}\rangle = 2H|t_l^j\rangle - |t_l^{j-1}\rangle.$$

We can see from the recursive relation that the computation of each column in the Chebyshev matrix $T_j(H)$ only requires matrix-vector multiplications. The computation cost of matrix-vector multiplications is $N_d \times n_H$, where n_H denotes the number of non-zero ele-

ments in each row of the Hamiltonian matrix. In general, the Hamiltonian matrix has a sparse representation, and the sparsity n_H is independent of the system size; hence the total computation cost of each column of the density matrix is proportional with $n_p \times N_d \times n_H$, where n_p is the degree of the Chebyshev polynomial approximation. The computation cost of the entire density matrix is then $N_d^2 \times n_H \times n_p$, which is an improvement over the cubic scaling cost for computing the density matrix exactly using diagonalization.

The computational cost can be further reduced if we can take advantage of the decay behavior of density matrix. We can define a localization region $|\mathbf{r} - \mathbf{r}'| \leq R_l$, which beyond R_l, the entry in the density operator $\gamma(\mathbf{r}, \mathbf{r}')$, is zero. Using the localized basis functions, if the distance between the center of the ith basis function and the jth basis function is beyond a distance $L(R_l)$, then $\gamma_{ij} = 0$. This means that for the kth column of the Chebyshev matrix, we only need to compute w_L elements above and below the kth element, where w_L depends only on R_L and n_H, independent of the system size. Then the computational cost of density matrix is proportional to $N_d \times w_L \times n_H \times n_p$, which has linear dependence on the system size.

In hind-sight, the Chebyshev method doesn't require truncation of the localization zone in order to achieve linear scaling, the computation cost of matrix-vector multiplications was just mis-estimated.

In order to arrive at the Chebyshev expansion of the Fermi-Dirac distribution, we need to know the Fermi energy λ_f. It can be found by using any root-finding algorithm that ensures that the trace of the density matrix equals the number of particles in the system.

Examples using Chebyshev approximations of the finite-temperature density matrix can be found in [22] and [6]. The degree of Chebyshev polynomials required for a given accuracy has been studied by Baer *et al.* in [6], and they have shown that the degree n_p for an accuracy 10^{-D} is

$$n_p \approx \frac{2}{3}(D-1)\beta_s, \tag{3.7}$$

where $\beta_s = \frac{\Delta\lambda}{2k_{\mathrm{B}}T}$, and $\Delta\lambda = \lambda_{\max} - \lambda_{\min}$ is the difference between the largest and the smallest

eigenvalues of the Hamiltonian matrix. Equation (3.7) will become important in section 3.0.4 later.

3.0.3.2 Linear scaling spectral Gauss quadrature

Another flavor of polynomial approximation of the density matrix is the approximation of the matrix trace using Gaussian quadratures along the spectrum of the Hamiltonian matrix [73]. This approximation is called the linear scaling spectral Gauss quadrature (LSSGQ) method. To illustrate the LSSGQ approximation, we should consider the expression of the Kohn-Sham total energy in equation (2.31) using the Kohn-Sham orbital energies from the eigenvalue problem in equation (2.29):

$$\epsilon_0(\{\mathbf{R_1}, \cdots, \mathbf{R_M}\}) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}d\mathbf{r}' - \int_{\mathbb{R}^3} \rho(\mathbf{r}) v_{\text{exc}}(\mathbf{r}) d\mathbf{r} + E_{\text{exc}}(\rho), \quad (3.8)$$

where λ_i corresponds to the ith lowest eigenvalues of the Kohn-Sham Hamiltonian matrix in equation (2.29), and $\rho(\mathbf{r})$ is the ground-state electron density. The first term in equation (3.8) can be written as simply

$$\sum_{i=1}^{N} \lambda_i = \text{Tr}(H\gamma). \quad (3.9)$$

Using the fact that the ground-state one-particle density matrix γ shares the same eigenstates as the Hamiltonian matrix, we can write the matrix trace in equation (3.9) as a summation of a family of spectral integrals along the spectrum of the Hamiltonian matrix:

$$\text{Tr}(H\gamma) = \sum_{i=1}^{\infty} \int_{\sigma(H)} \lambda g(\lambda) d\mu_{(\xi_i,\xi_i)}(\lambda), \quad (3.10)$$

where $g(\lambda)$ is the zero-temperature matrix function defined in equation (3.4), and $\mu_{(\xi_i,\xi_i)}$ is spectral measure defined by the projection of the resolution of identity $P(\lambda)$ associated with H onto the vector ξ_i:

$$\mu_{(\xi_i,\xi_i)} = \langle \xi_i | P(\lambda) | \xi_i \rangle.$$

$\{\xi_i\}$ is a set of complete orthonormal basis in $\mathcal{L}^2(\mathbb{R}^3)$.

The ground-state electron density at location $\mathbf{r}_0$ can also be written as a spectral integral:

$$\begin{aligned}
\rho(\mathbf{r}_0) &= \gamma(\mathbf{r}_0, \mathbf{r}_0) = \langle \mathbf{r}_0 | \gamma | \mathbf{r}_0 \rangle \\
&= \langle \mathbf{r}_0 | g(H) | \mathbf{r}_0 \rangle = \sum_{i=1}^{N} g(\lambda_i) \langle \mathbf{r}_0 | \psi_i \rangle \langle \psi_i | \mathbf{r}_0 \rangle \\
&= \sum_{i=1}^{N} g(\lambda_i) \langle \mathbf{r}_0 | \sum_{j=1}^{N_d} b_{ij} \xi_j \rangle \langle \sum_{k=1}^{N_d} b_{ik} \xi_k | \mathbf{r}_0 \rangle \\
&= \sum_{i=1}^{N} \sum_{j=1}^{N_d} \sum_{k=1}^{N_d} g(\lambda_i) b_{ij} b_{ik} \xi_j(\mathbf{r}_0) \xi_k(\mathbf{r}_0) \\
&= \sum_{j=1}^{N_d} \sum_{k=1}^{N_d} \sum_{i=1}^{N} g(\lambda_i) b_{ij} b_{ik} \xi_j(\mathbf{r}_0) \xi_k(\mathbf{r}_0) \\
&= \int_{\sigma(H)} g(\lambda) d\mu_{(\eta_{\mathbf{r}_0}, \eta_{\mathbf{r}_0})}(\lambda), \qquad (3.11)
\end{aligned}$$

where $|\psi_i\rangle = |\sum_{j=1}^{N_d} b_{ij} \xi_j\rangle$, $\eta_{\mathbf{r}_0} = \sum_{j=1}^{N_d} \xi_j(\mathbf{r}_0) \xi_j \rangle$, and $\mu_{(\eta_{\mathbf{r}_0}, \eta_{\mathbf{r}_0})}$ is the spectral measure defined by

$$\mu_{(\eta_{\mathbf{r}_0}, \eta_{\mathbf{r}_0})} = \langle \eta_{\mathbf{r}_0} | P(\lambda) | \eta_{\mathbf{r}_0} \rangle.$$

The LSSGQ approximation consists of approximating each of the spectral integrals in equation (3.10) using spectral Gauss quadratures. Since numerical quadratures are more efficient computationally when the integrands are smooth, so LSSGQ adopts the finite-temperature approximation of the density matrix, the Fermi-Dirac function in equation (3.6).

The key components of the spectral Gauss quadratures are the quadrature weights and nodes. Taking advantage of the sparsity of the Hamiltonian matrix, the computation of the spectral Gauss quadrature nodes and weights for each integral in equation (3.10) and equation (3.11) can be evaluated at $\mathcal{O}(1)$ cost, independent of the size of the system; resulting in a numerical scheme that scales linearly with respect to the system size for evaluation of $\mathcal{O}(N)$ spectral integrals.

To compute n_p spectral Gauss quadrature nodes and weights for the spectral integral with measure $\mu_{(\xi_i,\xi_i)}$, we construct the Krylov subspace of dimension n_p of H using ξ_i as the starting vector:

$$\mathcal{K}_{n_p,\xi_i} = \text{span}\{|\xi_i\rangle, H|\xi_i\rangle, \cdots, H^{n_p-1}|\xi_i\rangle\}. \tag{3.12}$$

The vectors in equation (3.12) are not orthonormal, so we can orthormalize them using the Lanczos method [39], which is a modified Gram-Schmidt orthogonalization procedure. The Lanczos algorithm works as follows, starting with $v^1\rangle = \xi_i\rangle, \alpha_i = \langle v^1|H|v^1\rangle, \tilde{v}^2\rangle = |Hv^1\rangle - |\alpha_1 v^1\rangle$, and then for $k = 2, 3, \cdots$:

$$\beta_{k-1} = \|\xi_i\|,$$

$$v^k = \frac{\tilde{v}^k}{\beta_{k-1}},$$

$$\alpha_k = \langle v^k|H|v^k\rangle,$$

$$\tilde{v}^{k+1}\rangle = |Hv^k\rangle - |\alpha_k v^k\rangle - \beta_{k-1}v^{k-1}.$$

We can collect the real numbers $\{\alpha_j\}$ and $\{\beta_j\}$ into a tridiagonal matrix J_{n_p}:

$$J_{n_p} = \begin{pmatrix} \alpha_1 & \beta_1 & & & \\ \beta_1 & \alpha_2 & \beta_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \beta_{n_p-2} & \alpha_{n_p-1} & \beta_{n_p-1} \\ & & & \beta_{n_p-1} & \alpha_{n_p} \end{pmatrix} \tag{3.13}$$

Let t_i and $d_i\rangle$ denote ith eigenvalue and eigenvector of the matrix J_{n_p} correspondingly. Then $\{t_i\}_{i=1}^{n_p}$ are the spectral Gauss quadrature nodes for the spectral integral with measure $\mu_{(\xi_i,\xi_i)}$, and the spectral Gauss quadrature weights are defined by

$$w_i = |d_i^1|^2,$$

where d_i^1 is the first element in the ith eigenvector. A detailed discussion of the spectral Gaussian quadrature nodes and weights can be found in [24]. In summary, the key distinction between Chebyshev polynomial approximation of the density matrix and the LSSGQ approximation is that in LSSGQ, we are not approximating the density matrix by a single polynomial function of the Hamiltonian matrix; it looks like a polynomial approximation because the spectral Gauss quadrature has close ties to polynomial approximations.

Suryanarayana [70] studied the convergence of the LSSGQ approximation with respect to a linear Hamiltonian matrix, and found that the rate of the convergence of the LSSGQ approximation scales proportionally to

$$\frac{2\pi\hat{\sigma}}{\sqrt{1-\hat{\lambda_f}^2}},$$

where $\hat{\sigma} = \frac{k_B T}{\Delta\lambda}$ and $\hat{\lambda}_f = \frac{\lambda_f}{\Delta\lambda}$. Similar to the Chebyshev polynomial approximation of the density matrix in section 3.0.3.1, the approximation error is proportional to the spectrum width $\Delta\lambda$ of the Hamiltonian matrix.

3.0.3.3 Rational approximation of density matrix

Goedecker [22] introduced a rational approximation of the density matrix using contour integration. The function $f(\lambda)$:

$$f(\lambda) = \frac{1}{2\pi i}\oint_C \frac{dz}{\lambda - z},$$

equals 1 if λ is enclosed by the contour C, and 0 otherwise. We can choose a contour that encloses exactly the occupied eigenvalues of the Hamiltonian matrix to approximate the zero-temperature density matrix. To apply to finite temperature, we can use any rational

functions that approximate the Fermi-Dirac distribution:

$$g_{\text{fermi}}(\lambda) \approx \sum_{i=1}^{n_r} \frac{w_i}{\lambda - t_i}.$$

As for the specific $\{t_i\}$ and $\{w_i\}$, Goedecker [22] used uniform spaced nodes on the contour curve; Lin *et al.* [45] applied fast multipole method to the Matsubara pole-expansion of the Fermi-Dirac function. To evaluate the rational expansion of the Fermi-Dirac function at linear-cost, we will need to evaluate projections of the inverse of matrices, $H - t_i I$. This is equivalent to solving linear systems of equations. We can use iterative methods like conjugate gradient methods. Lin *et al.* [45] proposed an algorithm for selected inversion of sparse symmetric matrix (Sellnv) involving LDL^T transform of the matrices $H - t_i I$, which is exact. However, the algorithm scales linearly with respect the number of electrons only in the system for quasi one dimensional systems; for three dimensional molecular systems, the algorithm scales quadratically with respect to the system size. Similar to the polynomial approximations, the number of rational functions required for a given accuracy scales inversely with respect to the temperature T in the Fermi-Dirac function, and proportionally to the spectrum width $\Delta\lambda$ of the Hamiltonian matrix. Lin *et al.* [45] showed that the number of poles required given an accuracy is,

$$n_r \propto ln(\frac{\Delta\lambda}{k_{\text{B}}T}).$$

3.0.4 The relationship between the spectrum width $\Delta\lambda$ and the system size

The number of expansions in the density matrix methods described above scale proportionally to the spectrum width $\Delta\lambda$ of the linearized Hamiltonian matrix, and the algorithm scale linearly with respect to the system size if and only if $\Delta\lambda$ is independent of the system size. In all the papers referenced above, the independence of $\Delta\lambda$ from the systems has not been

rigorously proven. In this section, we will show that for under certain assumptions, $\Delta\lambda$ is independent of the system size when the Hamiltonian is discretized using a central finite difference scheme.

Let $V_N(\mathbf{r}) = \sum_{i=1}^{N} v(\mathbf{r} - \mathbf{R_i})$ where R_i denotes the position of the nuclei and N denotes the number of atoms in the system. Assume that $v(\mathbf{r} - \mathbf{R_i}) \in L^\infty(\Omega)$ and decays faster than $\frac{1}{|\mathbf{r}-\mathbf{R_i}|}$ away from $\mathbf{R_i}$, such that we have $V_{\min} \leq V_N(\mathbf{r}) \leq V_{\max}$ independent of N. Consider the linear eigenvalue problem in Ω_N with periodic boundary condition:

$$H_{V_N}\psi(\mathbf{r}) = \{-\Delta + V_N(\mathbf{r})\}\psi(\mathbf{r}) = \lambda^{V_N}\psi(\mathbf{r}). \tag{3.14}$$

We show next that for central finite difference approximation with fixed discretization size Δr, the spectrum width $\Delta\lambda$ is independent of the system size N.

Consider two linear eigenvalue problems with a constant potential in the same domain as H_{V_N} with periodic boundary condition:

$$H_{V_{\min}}u(\mathbf{r}) = -\Delta + V_{\min}u(\mathbf{r}) = \lambda^{V_{\min}}u(\mathbf{r}) \tag{3.15}$$

and

$$H_{V_{\max}}u(\mathbf{r}) = -\Delta + V_{\max}u(\mathbf{r}) = \lambda^{V_{\max}}u(\mathbf{r}). \tag{3.16}$$

The proof can be further broken into 2 parts,

1. prove $\lambda_{\min}^{V_{\min}}(\Delta\mathbf{r}) \leq \lambda_{\min}^{V_N}(\Delta\mathbf{r}) < \lambda_{\max}^{V_N}(\Delta\mathbf{r}) \leq \lambda_{\max}^{V_{\max}}(\Delta\mathbf{r})$, where $\lambda_{\{\}}^{\{\}}(\Delta\mathbf{r})$ denotes the eigenvalues of the discretized Hamiltonian matrix $H^{\{\}}$.

2. prove the bound $\{\lambda_{\max}^{V_{\max}}(\Delta\mathbf{r}) - \lambda_{\min}^{V_{\min}}(\Delta\mathbf{r})\}$ is independent of N.

Part 1

Discretize $H^{V_{\min}}$ and $H^{V_{\max}}$ using the same central difference scheme with discretization size $\Delta\mathbf{r}$. Let's denote the discretized matrix of $H^{V_{\min}}$, H^{V_N}, $H^{V_{\max}}$ by $H^{(V_{\min},\Delta\mathbf{r})}$, $H^{(V_N,\Delta\mathbf{r})}$, $H^{(V_{\max},\Delta\mathbf{r})}$. Let $N_d(N)$ denotes the size of the matrix.

$H^{(V_{\min},\Delta\mathbf{r})}$, $H^{(V_N,\Delta\mathbf{r})}$, $H^{(V_{\max},\Delta\mathbf{r})}$ are real symmetric matrix with real eigenvalues. Let $\psi_{\max}$ denote the eigenvector corresponding the largest eigenvalue of the matrix $H^{(V_N,\Delta\mathbf{r})}$. Consider the quantity:

$$\begin{aligned}
&\langle\psi_{\max}|H^{(V_{\max},\Delta\mathbf{r})}|\psi_{\max}\rangle \\
&= \langle\psi_{\max}|[H^{V_N,\Delta\mathbf{r}} + H^{\left((V_{\max}-V_N),\Delta\mathbf{r}\right)}]|\psi_{\max}\rangle \\
&= \langle\psi_{\max}|H^{(V_N,\Delta\mathbf{r})}|\psi_{\max}\rangle + \langle\psi_{\max}|\underbrace{H^{\left((V_{\max}-V_N),\Delta\mathbf{r}\right)}}_{\text{positive semi-definite}}\psi_{\max}\rangle \\
&\geq \lambda_{\max}^{V_N}(\Delta\mathbf{r})
\end{aligned}$$

where $\langle\cdot,\cdot\rangle$ denotes the dot product between two vectors.

Now let $\{\xi_i\}$ denote the orthonormal eigenvectors of the matrix $H^{(V_{\max},\Delta\mathbf{r})}$, and we can expand the eigenvector of $H^{(V_N,\Delta\mathbf{r})}$: $\psi_{\max} = \sum_{i=1}^{N_d} c_i\xi_i$. By normalization, we have $\sum_{i=1}^{N_d} c_i^2 = 1$. Consider the same quantity:

$$\begin{aligned}
&\langle\psi_{\max}|H^{(V_{\max},\Delta\mathbf{r})}|\psi_{\max}\rangle \\
&= \langle\sum_{i=1}^{N_d} c_i\xi_i|H^{(V_{\max},\Delta\mathbf{r})}|\sum_{j=1}^{N_d} c_j\xi_j\rangle \\
&= \sum_{i=1}^{N_d}\lambda_i^{V_{\max}}c_i^2 \\
&\leq \lambda_{\max}^{V_{\max}}.
\end{aligned}$$

Hence we have shown that $\lambda_{\max}^{V_N}(\Delta\mathbf{r}) \leq \lambda_{\max}^{V_{\max}}(\Delta\mathbf{r})$.

Similarly, if we consider the product where $\psi_{\min}$ denotes the eigenvector corresponding

to the minimum eigenvalue of the matrix $H^{(V_N,\Delta\mathbf{r})}$.

$$\begin{aligned}
&\langle\psi_{\min}|H^{(V_{\min},\Delta\mathbf{r})}|\psi_{\min}\rangle \\
&= \langle\psi_{\min}|[H^{(V_N,\Delta\mathbf{r})} + H^{\left((V_N-V_{\min},\Delta\mathbf{r}\right)}]|\psi_{\min}\rangle \\
&= \langle\psi_{\min}|H^{(V_N,\Delta\mathbf{r})}|\psi_{\min}\rangle + \langle\psi_{\min}|\underbrace{H^{\left((V_{\min}-V_N),\Delta\mathbf{r}\right)}}_{\text{negative semi-definite}}|\psi_{\min}\rangle \\
&\leq \lambda_{\min}^{V_N}(\Delta\mathbf{r}).
\end{aligned}$$

Similarly, let $\{\xi_i\}$ denote the orthonormal eigenvectors of the matrix $H^{(V_{\min},\Delta\mathbf{r})}$, we can expand the eigenvector of $H^{(V_N,\Delta\mathbf{r})}$: $\psi_{\min} = \sum_{i=1}^{N_d} c_i\xi_i$. By normalization, we have $\sum_{i=1}^{N_d} c_i^2 = 1$. Consider the same quantity:

$$\begin{aligned}
&\langle\psi_{\min}|H^{(V_{\min},\Delta\mathbf{r})}|\psi_{\min}\rangle \\
&= \langle\sum_{i=1}^{N_d} c_i\xi_i|H^{(V_{\min},\Delta\mathbf{r})}|\sum_{j=1}^{N_d} c_j\xi_j\rangle \\
&= \sum_{i=1}^{N_d} \lambda_i^{V_{\min}} c_i^2 \\
&\geq \lambda_{\min}^{V_{\min}}
\end{aligned}$$

Hence we have shown that $\lambda_{\min}^{V_N}(\Delta\mathbf{r}) \geq \lambda_{\min}^{V_{\min}}(\Delta\mathbf{r})$.

Part 2 To obtain a bound for the gap $\Delta\lambda$, consider a one-dimensional infinite system subject to the Hamiltonian with a constant potential, as illustrated in equations (3.15) and (3.16). Depending on the order of the central difference scheme used, the discretized eigenvalue problem becomes

$$\cdots + c_2\psi_{n-2} + c_1\psi_{n-1} + c_0\psi_n + c_1\psi_{n+1} + c_2\psi_{n+2} + \cdots + V\psi_n = \lambda^{V,\Delta r}\psi_n, \tag{3.17}$$

where $\{c_i\}$ corresponds to the coefficient of the central difference scheme. We can make a

solution ansatz of the form,

$$\psi_n = A \exp(ikn\Delta r), \tag{3.18}$$

where k is the wavenumber. After substituting equation (3.18) into equation (3.17), we get the numerical dispersion relation for the discretized eigenvalue problem,

$$\lambda^{V,\Delta r}(k) = c_0 + 2c_1 \cos(k\Delta r) + 2c_2 \cos(2k\Delta r) + \cdots + V. \tag{3.19}$$

Although this is the dispersion relation for an infinite system, when we impose a boundary condition for a finite system, we are only limiting the largest wavelength (i.e., the smallest wavenumber k) of plane-waves the system can sustain. In the limit of $k \to 0$, the dispersion relation in equation (3.19) becomes:

$$\lambda^{V,\Delta r}(k) \approx c_0 + \sum_{n=1}^{p} 2c_n - \sum_{n=1}^{p} 2c_n \frac{(nk\Delta r)^2}{2} + V,$$

to a second order approximation of k. Since

$$c_0 + \sum_{i=1}^{p} 2c_i = 0,$$

for all central finite difference scheme for the Laplacian operator, at $k = 0$, we have

$$\lambda^{V,\Delta r}(k = 0) = V.$$

This condition dictates the lowest bound on the eigenvalues λ. We also notice that the numerical dispersion relation in equation (3.19) is simply a linear combination of cosine functions, hence there will be a maximum energy state $\lambda_{\max}$, and it will only be a function of Δr since it's the only parameter in equation (3.19). An intuitive way of thinking about this result is that since the potential energy is constant at V, the kinetic energy $|\nabla\psi|^2$ is what dictates the total energy λ; so the eigenvectors with a larger wave number k, the

larger its total energy λ. However, the discretization cannot support waves with arbitrarily large wave numbers. When the wavelength is shorter than $2\Delta r$, we will get aliasing effects. Figure 3.1 illustrates the numerical eigenvalues obtained from different orders of central difference schemes for a Hamiltonian with a constant potential. We can easily extended the numerical dispersion to a 3-dimensional system subject to a constant potential.

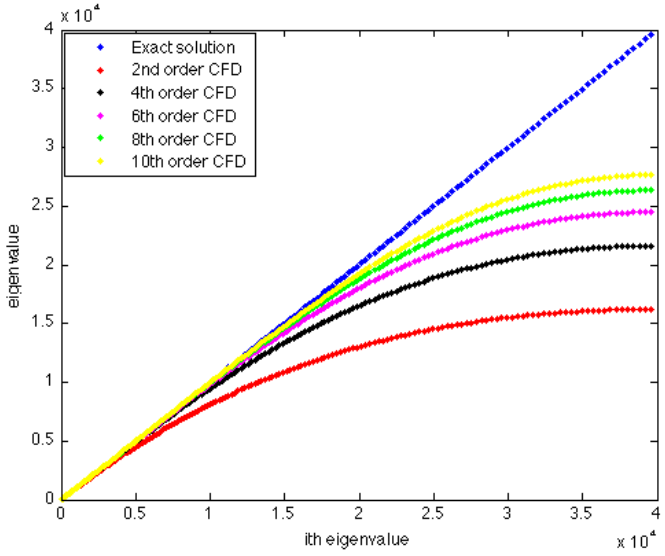

Figure 3.1: Illustration that the exact eigenvalues are always larger than the numerical eigenvalues.

Chapter 4

A variational frame work for spectral discretization in density functional theory

In the last chapter, we introduced linear-scaling density functional theory implementations that approximate the density matrix either by using polynomials or rational functions. Linear-scaling methods of this type, with or without truncation, often suffer from two significant shortcomings. Firstly, they approximate the density matrix of the linearized problem corresponding to an iteration of the self-consistent scheme; however, the global convergence properties of the entire self-consistent scheme itself, and of approximations thereof, are not well-established in general. Secondly, for reasons of computational expedience, linear-scaling methods often require severe smoothing of the occupancy function, corresponding to unphysically high temperatures.

In this chapter, we depart from the self-consistent scheme entirely and work directly with the variational formulation of KSDFT over trace-class operators. Anantharaman and Cancès [2] have used this variational formulation to prove the existence of solutions in bounded or unbounded domains. We use duality in the exchange-correlation functional to convert the classical variational formulation into nested variational problems. The resulting functional is linear in the density matrix and thus amenable to a simple spectral representation. Based on this reformulation, we introduce a new class of operator approximations, which we refer to as spectral binning. Spectral binning uses simple—or piecewise-constant—functions on

the spectrum and enables an accurate representation of the occupancy function without smoothing. The main mathematical result of this chapter consists of a proof of convergence of spectral binning with respect to combined spatial and spectral discretizations. As an example of application, we consider a standard one-dimensional benchmark problem (cf. [18]) and show that, for this problem, spectral binning exhibits excellent convergence characteristics and outperforms other linear-scaling methods.

The chapter is organized as follows: section 2 briefly reviews KSDFT 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.

4.1 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 and we hold the nuclei fixed throughout this section. We start with the operator formulation used by Anantharaman and Cancès, [2]. The connection to the traditional orbital formulation is given in Appendix B for completeness.

4.1.1 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 $\mathcal{H}$:

$$\mathfrak{S}_1 = \{\gamma \in \mathcal{S}(\mathcal{H}) : \mathrm{Tr}(|\gamma|) < \infty\}, \tag{4.1}$$

where $|\gamma| \equiv \sqrt{\gamma\gamma^*}$. $\mathfrak{S}_1$ is a separable Banach space [5]. Within $\mathfrak{S}_1$, we can introduce the space

$$\mathcal{X} = \{\gamma \in \mathfrak{S}_1 : |\nabla|\gamma|\nabla| \in \mathfrak{S}_1\}, \tag{4.2}$$

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

$$\mathcal{K}_N = \{\gamma \in \mathcal{X} \,:\, 0 \leq \gamma \leq 1, \mathrm{Tr}(\gamma) = N\}. \tag{4.3}$$

Remark 4.1.1 *As stated in [2], for every $\gamma \in \mathcal{K}_N$, we have the canonical representation in the continuous r basis,*

$$\gamma(\mathbf{r}, \mathbf{r}') = \sum_{i=1}^{\infty} 2\alpha_i \xi_i(\mathbf{r})\xi_i(\mathbf{r}'), \tag{4.4}$$

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

$$0 \leq \alpha_i \leq 1, \qquad \int_\Omega \xi_i(\mathbf{r})\xi_j(\mathbf{r})\,\mathrm{d}\mathbf{r} = \delta_{ij}, \qquad \sum_{i=1}^{\infty} 2\alpha_i = N. \tag{4.5}$$

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

$$\rho_\gamma(\mathbf{r}) = \gamma(\mathbf{r}, \mathbf{r}). \tag{4.6}$$

We consider a system of M atoms with nuclei located at $\{\mathbf{R_1}, \cdots, \mathbf{R_M}\} = \{\mathbf{R}_1, \ldots, \mathbf{R}_M\} \subset \Omega$ and nuclear charges $Z_1, \ldots, Z_M$. 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

$$E^{\mathrm{EKS}}(\gamma) = T_0(\gamma) + E_{\mathrm{H}}(\rho_\gamma) + E_{\mathrm{ext}}(\rho_\gamma) + U_{\mathrm{n-n}} + E_{\mathrm{xc}}(\rho_\gamma), \tag{4.7}$$

where T_0 is the kinetic energy of the non-interacting electrons,

$$T_0(\gamma) = \mathrm{Tr}\left(-\frac{1}{2}\Delta\gamma\right), \tag{4.8}$$

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

$$E_{\mathrm{H}}(\rho_\gamma) = \frac{1}{2}\int_\Omega\int_\Omega \frac{\rho_\gamma(\mathbf{r})\rho_\gamma(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{r}', \tag{4.9}$$

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

$$E_{\mathrm{ext}}(\rho_\gamma) = \int_\Omega \rho_\gamma(\mathbf{r}) v_{\mathrm{ext}}(\mathbf{r}, \{\mathbf{R_1}, \cdots, \mathbf{R_M}\})\,\mathrm{d}\mathbf{r} = \int_\Omega \rho_\gamma(\mathbf{r}) \left(\sum_{1\le I\le M} \frac{Z_I}{|\mathbf{R}_I - \mathbf{r}|}\right)\,\mathrm{d}\mathbf{r}, \tag{4.10}$$

$U_{\mathrm{n-n}}$ is the classical electrostatic repulsion energy due to the nuclear charges,

$$U_{\mathrm{n-n}} = \frac{1}{2}\sum_{1\le I\le J\le M} \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|}, \tag{4.11}$$

and $E_{\mathrm{xc}}(\rho_\gamma)$ is the exchange-correlation energy that is split into two terms (cf. [59]),

$$E_{\mathrm{xc}}(\rho_\gamma) = E_{\mathrm{x}}(\rho_\gamma) + E_{\mathrm{c}}(\rho_\gamma) = \int_\Omega h(\rho_\gamma)\,\mathrm{d}\mathbf{r}, \tag{4.12}$$

with an exchange term,

$$E_{\mathrm{x}}(\rho_\gamma) = -\frac{3}{4}\Big(\frac{6}{\pi}\Big)^{1/3}\int_\Omega \rho_\gamma^{4/3}(\mathbf{r})\,\mathrm{d}\mathbf{r}, \tag{4.13}$$

and a correlation term,

$$E_{\mathrm{c}}(\rho_\gamma) = \int_\Omega \epsilon_{\mathrm{c}}(\rho_\gamma(\mathbf{r}))\rho_\gamma(\mathbf{r})\,\mathrm{d}\mathbf{r}, \tag{4.14}$$

where ϵ_{c} is taken from [59]. 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

$$\epsilon_0^{\mathrm{EKS}} = \inf_{\gamma\in\mathcal{K}_N} E^{\mathrm{EKS}}(\gamma). \tag{4.15}$$

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

4.1.2 Reformulation

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

4.1.2.1 Electrostatics

We reformulate the electrostatic terms by writing them as the solution to a Helmholtz problem (cf., e.g., [35, 74]). 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

$$\int_\Omega f_{\mathbf{R}_i}(\mathbf{r})\,\mathrm{d}\mathbf{r} = 1. \tag{4.16}$$

We can then rewrite the electrostatic terms as the variational problem

$$\begin{aligned} &E_{\mathrm{H}}(\rho_\gamma) + E_{\mathrm{ext}}(\rho_\gamma) + U_{\mathrm{n-n}} \\ &\quad = \sup_{\phi\in\mathcal{V}} \left\{ -C_S \int_\Omega |\nabla\phi(\mathbf{r})|^2\,\mathrm{d}\mathbf{r} + \int_\Omega (b(\mathbf{r},\{\mathbf{R_1},\cdots,\mathbf{R_M}\}) + \rho_\gamma(\mathbf{r}))\phi(\mathbf{r})\,\mathrm{d}\mathbf{r} \right\} + C_{\mathrm{self}}, \end{aligned}$$

where

$$b(\mathbf{r},\{\mathbf{R_1},\cdots,\mathbf{R_M}\}) = \sum_{i=1}^{M} Z_i f_{\mathbf{R}_i}(\mathbf{r}). \tag{4.17}$$

$C_S > 0$ 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 $\{\mathbf{R_1},\cdots,\mathbf{R_M}\}$.

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

$$\Phi(\mathbf{r},\mathbf{r}') = \phi(\mathbf{r})\delta(\mathbf{r},\mathbf{r}'), \tag{4.18}$$

and use its coordinate representation so that

$$\mathrm{Tr}(\Phi\gamma) = \int_{\Omega} \phi(\mathbf{r})\rho_{\gamma}(\mathbf{r})\,\mathrm{d}\mathbf{r}. \tag{4.19}$$

The Coulomb energy is

$$\begin{aligned} J(\rho_{\gamma}) &= E_{\mathrm{H}}(\rho_{\gamma}) + E_{\mathrm{ext}}(\rho_{\gamma}) + U_{\mathrm{n-n}} \\ &= \sup_{\phi\in\mathcal{V}} \left\{ \mathrm{Tr}(\Phi\gamma) - C_S \int_{\Omega} |\nabla\phi(\mathbf{r})|^2\,\mathrm{d}\mathbf{r} + \int_{\Omega} b(\mathbf{r}, \{\mathbf{R_1}, \cdots, \mathbf{R_M}\})\phi(\mathbf{r})\,\mathrm{d}\mathbf{r} \right\} + C_{\mathrm{self}}. \end{aligned} \tag{4.20}$$

4.1.2.2 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 (4.12):

(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*:

$$h(0) = 0. \tag{4.21}$$

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

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

$$h'(t) \leq 0. \tag{4.22}$$

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

$$C_1|t|^{4/3} + C_2 \leq |h(t)| \leq C_3|t|^{4/3} + C_4, \tag{4.23}$$

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 4.1.2 *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 (4.23), with Fatou's Lemma it follows that $E_{\rm xc}(\rho_\gamma)$ 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

$$B_{\rm xc}(\rho_\gamma) = -E_{\rm xc}(\rho_\gamma). \tag{4.24}$$

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

$$\mathcal{U} = \mathcal{L}^4(\Omega) \tag{4.25}$$

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

$$B_{\rm xc}(\rho_\gamma) = \sup_{u \in \mathcal{U}} \{\langle \rho_\gamma, u\rangle - B^*_{\rm xc}(u)\}, \tag{4.26}$$

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

$$\langle v, u\rangle = \int_\Omega v(\mathbf{r})u(\mathbf{r})\,\mathrm{d}\mathbf{r}. \tag{4.27}$$

Using arguments from [17], we can rewrite the exchange-correlation functional,

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

Finally, we introduce the unbounded local operator

$$U(\mathbf{r}, \mathbf{r}') = u(\mathbf{r})\delta(\mathbf{r}, \mathbf{r}'), \tag{4.28}$$

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

$$E_{\mathrm{xc}}(\rho_\gamma) = \inf_{u\in\mathcal{U}}\{-\mathrm{Tr}(U\gamma) + B^*_{\mathrm{xc}}(u)\}. \tag{4.29}$$

4.1.2.3 Reformulated Extended Kohn-Sham Functional

Substituting (4.20) and (4.29) in (4.7) and omitting the inessential constant C_{self} for brevity, we obtain the reformulated extended KS(REKS) energy functional $E^{\mathrm{REKS}} : \mathcal{K}_N \to \mathbb{R}$ as

$$E^{\mathrm{REKS}}(\gamma) = \inf_{u\in\mathcal{U}} \sup_{\phi\in\mathcal{V}} L(u, \phi, \gamma), \tag{4.30}$$

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

$$L(u, \phi, \gamma) = \mathrm{Tr}(H(\phi, u)\gamma) + \int_\Omega \big(- C_S|\nabla\phi(\mathbf{r})|^2 + b(\mathbf{R}, \mathbf{r})\phi(\mathbf{r})\big)\,\mathrm{d}\mathbf{r} + B^*_{\mathrm{xc}}(u), \tag{4.31}$$

with the Hamiltonian

$$H(\phi, u) = -\frac{1}{2}\Delta + \Phi - U \tag{4.32}$$

and Φ, U defined in (4.18), (4.28). The ground-state energy of the system with M atoms is

$$\begin{aligned}
\epsilon_0^{\mathrm{REKS}} &= \inf_{\gamma\in\mathcal{K}_N} E^{\mathrm{REKS}}(\gamma) \\
&= \inf_{\gamma\in\mathcal{K}_N} \inf_{u\in\mathcal{U}} \sup_{\phi\in\mathcal{V}} L(u, \phi, \gamma) \qquad (4.33) \\
&= \inf_{\gamma\in\mathcal{K}_N} \inf_{u\in\mathcal{U}} \sup_{\phi\in\mathcal{V}} \left\{ \mathrm{Tr}(H(\phi, u)\gamma) + \int_\Omega \big(- C_S|\nabla\phi(\mathbf{r})|^2 + b(\mathbf{r}, \{\mathbf{R_1}, \cdots, \mathbf{R_M}\})\phi(\mathbf{r})\big)\,\mathrm{d}\mathbf{r} + B^*_{\mathrm{xc}}(u) \right\}.
\end{aligned}$$

4.2 Main results

We prove the following theorems on the reformulated extended KS functional.

Theorem 3 *The reformulated extended KS energy functional $E^{REKS}(\gamma)$ in (4.30) possesses a minimizer in $\mathcal{K}_N$.*

Theorem 4 *The order of the infimum and supremum in the computation of the ground-state energy of the reformulated KS energy functional (4.33) can be exchanged:*

$$\begin{aligned} \epsilon_0^{\mathrm{REKS}} &= \inf_{\gamma\in\mathcal{K}_N}\inf_{u\in\mathcal{U}}\sup_{\phi\in\mathcal{V}} L(u,\phi,\gamma) \\ &= \inf_{u\in\mathcal{U}}\sup_{\phi\in\mathcal{V}}\inf_{\gamma\in\mathcal{K}_N} L(u,\phi,\gamma), \end{aligned} \tag{4.34}$$

where L is given by (4.31).

Theorem 4 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 $\phi \in \mathcal{V}$,

$$\inf_{\gamma\in\mathcal{K}_N} \mathrm{Tr}(H(\phi,u)\gamma) \tag{4.35}$$

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 (4.83) below). We obtain a spectral discretization by limiting γ to $\mathcal{K}_{N,k}^H$ made of k simple functions of H (see (4.103) 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 5 *Let $k_j \to \infty$ as $j \to \infty$. Then, the diagonal sequence of spatially and spectrally discrete reformulated extended KS energies converges to the full KS ground-state energy:*

$$\lim_{j\to\infty} \inf_{\mathcal{U}_j} \sup_{\mathcal{V}_j} \inf_{\mathcal{K}_{N,k_j}^{Hj(\phi,u)}} L^j(u,\phi,\gamma) = \inf_{\mathcal{U}} \sup_{\mathcal{V}} \inf_{\mathcal{K}_N^{H(\phi,u)}} L(u,\phi,\gamma) = \epsilon_0^{\mathrm{REKS}}. \tag{4.36}$$

4.3 Existence of solutions

To establish the existence of minimizers in $\mathcal{K}_N$ for the KS-DFT problem in equation (4.30), we use tools similar to those used in 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. Specifically, we consider the weak*-topology of the vector space $\mathcal{X}$ endowed with the norm

$$\|\cdot\|_{\mathcal{X}} = \mathrm{Tr}(|\cdot|) + \mathrm{Tr}(||\nabla|\cdot|\nabla||) \tag{4.37}$$

in the convex set $\mathcal{K}_N$ defined in (4.3).

For clarity of notation, in the remainder of this chapter, we change our notation on the repulsive energy functionals (4.12) and (4.20) in order to emphasize their dependence on the reduced one-particle density operator and write

$$E_{\mathrm{xc}}(\gamma) \equiv E_{\mathrm{xc}}(\rho_\gamma), \qquad J(\gamma) \equiv J(\rho_\gamma). \tag{4.38}$$

Remark 4.3.1 *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 4.3.1, see for instance Part II of Theorem 2.2.1 in [36].

Lemma 4.3.2 *For all $\gamma \in \mathcal{K}_N$, the following inequalities hold.*

1. Lower bound on the kinetic energy,

$$\frac{1}{2}\|\nabla\sqrt{\rho_\gamma}\| \leq \mathrm{Tr}(-\frac{1}{2}\Delta\gamma) = \frac{1}{2}\mathrm{Tr}(|\nabla|\gamma|\nabla|). \tag{4.39}$$

2. Lower bound on the Coulomb energy,

$$0 \leq J(\gamma). \tag{4.40}$$

3. Lower bound on the exchange-correlation energy,

$$-C_3|\Omega|^{-1/3}N^{4/3} - C_4|\Omega| \leq E_{\mathrm{xc}}(\gamma). \tag{4.41}$$

4. Lower bound on the reformulated extended KS energy functional,

$$\|\gamma\|_{\mathcal{X}} - C_5 \leq E^{\mathrm{REKS}}(\gamma) \tag{4.42}$$

for a constant $C_5 > 0$ independent of γ. In particular, by (4.42), $E^{\mathrm{REKS}}(\gamma)$ 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

$$\rho_\gamma(\mathbf{r}) = \sum_{i=1}^{\infty} 2\alpha_i\xi_i(\mathbf{r})^2. \tag{4.43}$$

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

$$\frac{1}{2}\|\nabla\sqrt{\rho_\gamma}\|_{\mathcal{L}^2(\Omega)} \leq \mathrm{Tr}(-\frac{1}{2}\Delta\gamma) = \frac{1}{2}\mathrm{Tr}(|\nabla|\gamma|\nabla|). \tag{4.44}$$

2. *Lower bound on the Coulomb energy.* We have

$$J(\gamma) = \sup_{\phi\in\mathcal{V}}\left\{\int_\Omega \phi(\mathbf{r})(b(\{\mathbf{R_1},\cdots,\mathbf{R_M}\},\mathbf{r}) + \rho_\gamma(\mathbf{r}))\,\mathrm{d}\mathbf{r} - C_S\int_\Omega |\nabla\phi(\mathbf{r})|^2\,\mathrm{d}\mathbf{r}\right\} \geq 0, \tag{4.45}$$

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 (B.6) in Appendix B, the LDA exchange-correlation functional integrand h in equation (4.12) is bounded from below:

$$\begin{aligned}
E_{\mathrm{xc}}(\gamma) &= \inf_{u\in\mathcal{U}}\{-\mathrm{Tr}(U\gamma) + B^*_{\mathrm{xc}}(u)\} \\
&\geq \inf_{u\in\mathcal{U}}\{-\mathrm{Tr}(U\gamma) + C_{18}\|u\|^4_{\mathcal{L}^4(\Omega)} + C_{19}|\Omega|\} \\
&= -\mathrm{Tr}(U_\gamma\gamma) + C_{18}\|u_\gamma\|^4_{\mathcal{L}^4(\Omega)} + C_{19}|\Omega| \qquad (4.46)\\
&\geq -\mathrm{Tr}(U_\gamma\gamma) + C_{19}|\Omega| \\
&\geq -C|\Omega|^{-1/3}\big(\mathrm{Tr}(\gamma)\big)^{4/3} + C_{19}|\Omega| \\
&= -C|\Omega|^{-1/3}N^{4/3} + C_{19}|\Omega|, \qquad (4.47)
\end{aligned}$$

where u_γ denotes a minimizer of equation (4.46) and U_γ is its corresponding operator. It is evident that there exists a minimizer for the variational problem (4.46).

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

Putting together all the inequalities in the equations (4.45) and (4.47), we end up with

$$E^{\mathrm{REKS}}(\gamma) \geq \mathrm{Tr}\Big(-\frac{1}{2}\Delta\gamma\Big) - C|\Omega|^{-1/3}N^{4/3} + C_{19}|\Omega| = \frac{1}{2}\big(\mathrm{Tr}(|\nabla|\gamma|\nabla|) + \mathrm{Tr}(|\gamma|)\big) - C_5. \tag{4.48}$$

Here, we introduced the new constant

$$C_5 \equiv C|\Omega|^{-1/3}N^{4/3} - C_{19}|\Omega| + \frac{N}{2}. \tag{4.49}$$

For the derivation of (4.48), we used that for every $\gamma \in \mathcal{K}_N$, directly from the definition of this set,

$$\mathrm{Tr}(\gamma) = \mathrm{Tr}(|\gamma|) = N. \tag{4.50}$$

The estimate (4.48) implies that for any $t \in \mathbb{R}$ the level sets

$$\big\{\gamma \in \mathcal{K}_N : E^{\mathrm{REKS}}(\gamma) \leq t\big\} \tag{4.51}$$

are bounded,

$$t + C_5 \geq \frac{1}{2}(\mathrm{Tr}(|\gamma|) + \mathrm{Tr}(|\nabla|\gamma|\nabla|)) \equiv \frac{1}{2}\|\gamma\|_{\mathcal{X}}. \tag{4.52}$$

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 4.3.3 *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

$$W = |\psi\rangle\langle\psi|, \tag{4.53}$$

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

$$0 \leq \lim_{n\to\infty} \mathrm{Tr}(\gamma_n W) = \mathrm{Tr}(\gamma W), \tag{4.54}$$

and

$$\mathrm{Tr}(\gamma W) = \lim_{n\to\infty} \mathrm{Tr}(\gamma_n W) = \lim_{n\to\infty} \langle \psi, \gamma_n \psi \rangle \leq \langle \psi, \psi \rangle = 1. \tag{4.55}$$

Since the estimate (4.55) holds for all normalized $\psi \in \mathcal{H}$, we find with (4.54) 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 [11]. From equation (4.39) 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

$$\lim_{n\to\infty} \mathrm{Tr}(\gamma_n) = \lim_{n\to\infty} \int_\Omega \rho_{\gamma_n}\, \mathrm{d}\mathbf{r} = \lim_{n\to\infty} \|\sqrt{\rho_{\gamma_n}}\|_{\mathcal{L}^2}^2 = \|\sqrt{\rho_\gamma}\|_{\mathcal{L}^2(\Omega)}^2 = \int_\Omega \rho_\gamma\, \mathrm{d}\mathbf{r} = \mathrm{Tr}(\gamma). \tag{4.56}$$

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

Lemma 4.3.4 *The functional $J(\gamma)$ introduced in (4.20) 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 $\mathcal{K}_N$:

$$\begin{aligned}
|\mathrm{Tr}(\Phi\gamma)| &= \Big| \sum_{i=1}^{\infty} \langle \Phi\gamma\xi_i, \xi_i\rangle \Big| \leq \sum_{i=1}^{\infty} 2\alpha_i |\langle \Phi\xi_i, \xi_i\rangle| \\
&\leq \sum_{i=1}^{\infty} 2\alpha_i \|\phi\|_{\mathcal{L}^2(\Omega)} \|\xi_i^2\|_{\mathcal{L}^2(\Omega)} = \|\phi\|_{\mathcal{L}^2(\Omega)} \sum_{i=1}^{\infty} 2\alpha_i \|\xi_i\|_{\mathcal{L}^4(\Omega)}^2 \\
&\leq C\|\phi\|_{\mathcal{L}^2(\Omega)} \sum_{i=1}^{\infty} 2\alpha_i \|\nabla\xi_i\|_{\mathcal{L}^2(\Omega)}^2 = C\|\phi\|_{\mathcal{L}^2(\Omega)} \mathrm{Tr}(-\Delta\gamma),
\end{aligned} \tag{4.57}$$

where $\{\xi_i\}_{i\in\mathbb{N}}$ come from the canonical representation of $\gamma \in \mathcal{K}_N$, cf. equation (4.4), and the Gagliardo–Nirenberg–Sobolev inequality has been used to obtain equation (4.57). Consequently,

$$J(\gamma) = \sup_{\phi\in\mathcal{V}} \left\{ \mathrm{Tr}(\Phi\gamma) + \int_\Omega \big(b(\mathbf{r}, \{\mathbf{R_1}, \cdots, \mathbf{R_M}\})\phi(\mathbf{r}) - C_S|\nabla\phi(\mathbf{r})|^2 \big) d\mathbf{r} \right\} \tag{4.58}$$

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

Proof Similarly to the proof of Lemma 4.3.4, 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

$$-\mathrm{Tr}(U\gamma_n) + B^*_{\mathrm{xc}}(u). \tag{4.59}$$

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

$$-\mathrm{Tr}(U\gamma) + B^*_{\mathrm{xc}}(u) \tag{4.60}$$

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 $u_n \rightharpoonup u$,

$$\liminf_{n\to\infty} \{-\mathrm{Tr}(U_n\gamma_n) + B^*_{\mathrm{xc}}(u_n)\} \geq -\mathrm{Tr}(U\gamma) + B^*_{\mathrm{xc}}(u). \tag{4.61}$$

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

$$\lim_{n\to\infty} \langle \gamma_n\xi_i, v\rangle = \langle \gamma\xi_i, v\rangle. \tag{4.62}$$

From the proof of Lemma 4.3.3, 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. This completes the proof of the lim-inf condition.

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

implying

$$\limsup_{n\to\infty}\{-\mathrm{Tr}(U_n\gamma_n) + B^*_{\mathrm{xc}}(u_n)\} \geq -\mathrm{Tr}(U\gamma) + B^*_{\mathrm{xc}}(u). \tag{4.63}$$

Lastly, to show equi-coercivity of the functionals, from equation (B.6) in Appendix B,

$$-\mathrm{Tr}(u\gamma_n) + B^*_{\mathrm{xc}}(u) \geq C_{18}\|u\|^4_{\mathcal{U}} - (\sup_n C_n)\|u\|_{\mathcal{L}^2(\Omega)} + C_{19}|\Omega|, \tag{4.64}$$

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

$$-\mathrm{Tr}(u\gamma_n) + B^*_{\mathrm{xc}}(u) \tag{4.65}$$

is equi-coercive. Using Theorem 7.8 in [49], we have

$$\lim_{n\to\infty} E_{\mathrm{xc}}(\gamma_n) = \lim_{n\to\infty}\inf_{u\in\mathcal{U}}\{-\mathrm{Tr}(U\gamma_n) + B^*_{\mathrm{xc}}(u)\} = \inf_{u\in\mathcal{U}}\{\mathrm{Tr}(U\gamma) + B^*_{\mathrm{xc}}(u)\} = E_{\mathrm{xc}}(\gamma). \qquad \blacksquare \tag{4.66}$$

Lemma 4.3.6 *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*

$$E^{\mathrm{REKS}}(\gamma) \leq \liminf_{n\to\infty} E^{\mathrm{REKS}}(\gamma_n). \tag{4.67}$$

Proof To prove the lower semi-continuity of $E^{\mathrm{REKS}}(\gamma)$, we use the continuity of the functional $J(\gamma)$ from Lemma 4.3.4 and the continuity of $E_{\mathrm{xc}}(\gamma)$ from Remark 4.1.2 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

$$\begin{aligned}
\mathrm{Tr}(-\Delta\gamma) &= \mathrm{Tr}(|\nabla|\gamma|\nabla|) \\
&= \sum_{k=1}^{\infty} \langle\psi_k\big||\nabla|\gamma|\nabla|\big|\psi_k\rangle \\
&= \sum_{k=1}^{\infty} \mathrm{Tr}\big(\gamma(\big||\nabla|\psi_k\rangle\langle|\nabla|\psi_k\big|)\big) \\
&= \sum_{k=1}^{\infty} \lim_{n\to\infty} \mathrm{Tr}\big(\gamma_n(\big||\nabla|\psi_k\rangle\langle|\nabla|\psi_k\big|)\big) \\
&\leq \liminf_{n\to\infty} \sum_{k=1}^{\infty} \mathrm{Tr}\big(\gamma_n(\big||\nabla|\psi_k\rangle\langle|\nabla|\psi_k\big|)\big) \\
&= \liminf_{n\to\infty} \mathrm{Tr}(|\nabla|\gamma_n|\nabla|). \qquad (4.68)
\end{aligned}$$

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

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

Proof Consider a minimizing sequence $\{\gamma_n\}_{n\in\mathbb{N}}$ of $E^{\mathrm{REKS}}(\gamma)$ in $\mathcal{K}_N$. From Lemma 4.3.2 and Lemma 4.3.1, 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

$$\inf_{\gamma\in\mathcal{K}_N} E^{\mathrm{REKS}}(\gamma) \leq E^{\mathrm{REKS}}(\gamma_0) \leq \liminf_{n\to\infty} E^{\mathrm{REKS}}(\gamma_n) = \inf_{\gamma\in\mathcal{K}_N} E^{\mathrm{REKS}}(\gamma). \qquad (4.69)$$

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

4.4 Discretization of the energy functional

Next, we introduce the spectral and spatial discretizations of the reformulated extended KS functional and prove the convergence of the resulting approximate solutions.

4.4.1 Justification of the spectral discretization

Before we can apply spectral discretization, as it will become evident subsequently, we need 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 (4.31):

$$L(u, \phi, \gamma) = \mathrm{Tr}(H(\phi, u)\gamma) + \int_\Omega \left(-C_S|\nabla\phi(\mathbf{r})|^2 + b(\{\mathbf{R_1}, \cdots, \mathbf{R_M}\}, \mathbf{r})\phi(\mathbf{r})\right) \mathrm{d}\mathbf{r} + B^*_{\mathrm{xc}}(u). \quad (4.70)$$

The ground-state energy equals, cf. the equations (4.30) and (4.31),

$$\epsilon_0^{\mathrm{REKS}} = \inf_{\gamma\in\mathcal{K}_N} \inf_{u\in\mathcal{U}} \sup_{\phi\in\mathcal{V}} L(u, \phi, \gamma). \quad (4.71)$$

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

$$\epsilon_0^{\mathrm{REKS}} = \inf_{u\in\mathcal{U}} \inf_{\gamma\in\mathcal{K}_N} \sup_{\phi\in\mathcal{V}} L(u, \phi, \gamma). \quad (4.72)$$

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 4.4.1 *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 $\phi \in \mathcal{V}$,*

$$\lim_{\|\gamma\|_{\mathcal{X}}\to+\infty} L(u, \phi, \gamma) = +\infty. \quad (4.73)$$

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 4.3.6 we observe that $\mathrm{Tr}(-\frac{1}{2}\Delta\gamma)$ is lower semi-continuous in $\mathcal{X}$. Since, for every sequence $\gamma_n \to \gamma$ in $\mathcal{K}_N$, by compact embedding $\rho_{\gamma_n} \to \rho_\gamma$ in $\mathcal{L}^2(\Omega)$, the functionals $\mathrm{Tr}(\Phi\gamma)$ and $\mathrm{Tr}(U\gamma)$ are also continuous

in $\mathcal{X}$.

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

$$\begin{aligned} L(u,\phi,\gamma) &= \mathrm{Tr}(-\frac{1}{2}\Delta\gamma) + \mathrm{Tr}(\Phi\gamma) - \mathrm{Tr}(U\gamma) \\ &\geq \mathrm{Tr}(-\frac{1}{2}\Delta\gamma) - (\|u\|_{\mathcal{L}^2(\Omega)} + \|\phi\|_{\mathcal{L}^2(\Omega)})\|\rho_\gamma\|_{\mathcal{L}^2(\Omega)} \\ &\geq \mathrm{Tr}(-\frac{1}{2}\Delta\gamma) - C_6(\|u\|_{\mathcal{L}^2(\Omega)} + \|\phi\|_{\mathcal{L}^2(\Omega)})\|\rho_\gamma\|^{\frac{1}{4}}_{\mathcal{L}^1(\Omega)}\|\rho_\gamma\|^{\frac{3}{4}}_{\mathcal{L}^3(\Omega)} \qquad (4.74) \\ &\geq \mathrm{Tr}(-\frac{1}{2}\Delta\gamma) - C_7(\|u\|_{\mathcal{L}^2(\Omega)} + \|\phi\|_{\mathcal{L}^2(\Omega)})\mathrm{Tr}(|\gamma|)^{\frac{1}{4}}\|\nabla\sqrt{\rho_\gamma}\|^{\frac{3}{2}}_{\mathcal{L}^2(\Omega)} \qquad (4.75) \end{aligned}$$

for some positive real constants C_6 and C_7, where interpolation inequalities are used to obtain equation (4.74) and the Gagliardo–Nirenberg–Sobolev inequality is used to obtain equation (4.75). Hence

$$L(u,\phi,\gamma) \geq \frac{1}{2}\|\gamma\|_{\mathcal{X}} - C_8\|\nabla\sqrt{\rho_\gamma}\|^{\frac{3}{2}}_{\mathcal{L}^2(\Omega)} - \frac{N}{2}, \qquad (4.76)$$

where $C_8 \equiv C_7 N^{1/4}(\|u\|_{\mathcal{L}^2(\Omega)} + \|\phi\|_{\mathcal{L}^2(\Omega)})$, implying the coercivity (4.73) of $L(u,\phi,\cdot)$. ∎

Lemma 4.4.2 *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,*

$$\lim_{\|\phi\|_{\mathcal{V}} \to +\infty} L(u,\phi,\gamma) = -\infty. \qquad (4.77)$$

Proof For given u and γ, the terms $\mathrm{Tr}(\Phi\gamma)$ and $\int_\Omega b(\mathbf{r},\{\mathbf{R_1},\cdots,\mathbf{R_M}\})\phi(\mathbf{r})\,\mathrm{d}\mathbf{r}$ 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 4.4.1, we observe that $\mathrm{Tr}(\Phi\gamma)$ and $\int_\Omega b(\mathbf{r},\{\mathbf{R_1},\cdots,\mathbf{R_M}\})\phi(\mathbf{r})\,\mathrm{d}\mathbf{r}$ are continuous in $\mathcal{V}$ for given $b(\mathbf{r},\{\mathbf{R_1},\cdots,\mathbf{R_M}\})$ 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 [49].

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

$$\begin{aligned} -L(u,\phi,\gamma) &\geq C_S\|\nabla\phi\|^2_{\mathcal{L}^2(\Omega)} - \|\phi\|_{\mathcal{L}^2(\Omega)}\|\rho_\gamma + b(\mathbf{r},\{\mathbf{R_1},\cdots,\mathbf{R_M}\})\|_{\mathcal{L}^2(\Omega)} + C_9(u,\gamma) \\ &\geq C_{10}\|\phi\|^2_{\mathcal{L}^2(\Omega)} - \|\phi\|_{\mathcal{L}^2(\Omega)}\|\rho_\gamma + b(\mathbf{r},\{\mathbf{R_1},\cdots,\mathbf{R_M}\})\|_{\mathcal{L}^2(\Omega)} + C_9(u,\gamma), \end{aligned} \tag{4.78}$$

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

$$C_9(u,\gamma) \equiv \mathrm{Tr}(\frac{1}{2}\Delta\gamma) + \mathrm{Tr}(U\gamma) - B^*_{\mathrm{xc}}(u). \tag{4.79}$$

Applying Young's inequality to $\|\phi\|_{\mathcal{L}^2(\Omega)}\|\rho_\gamma + b(\mathbf{r},\{\mathbf{R_1},\cdots,\mathbf{R_M}\})\|_{\mathcal{L}^2(\Omega)}$ in (4.78), $\|\phi\|_{\mathcal{L}^2(\Omega)}$ 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 show that it is possible to exchange the orders of the infima and supremum when computing $\epsilon_0^{\mathrm{REKS}}$. This commutativity property is important, as it allows to apply spectral theory to the Lagrange functional $L(u,\phi,\gamma)$.

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

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

$$\begin{aligned} \epsilon_0^{\mathrm{REKS}} &= \inf_{\gamma\in\mathcal{K}_N}\inf_{u\in\mathcal{U}}\sup_{\phi\in\mathcal{V}} L(u,\phi,\gamma) \\ &= \inf_{u\in\mathcal{U}}\sup_{\phi\in\mathcal{V}}\inf_{\gamma\in\mathcal{K}_N} L(u,\phi,\gamma) \\ &= \inf_{u\in\mathcal{U}}\sup_{\phi\in\mathcal{V}}\inf_{\gamma\in\mathcal{K}_N}\left\{ E_{\mathrm{band}}(u,\phi,\gamma) + \int_\Omega \big(- C_S|\nabla\phi(\mathbf{r})|^2 + b(\mathbf{R},\mathbf{r})\phi(\mathbf{r})\big)\,\mathrm{d}\mathbf{r} + B^*_{\mathrm{xc}}(u)\right\}. \end{aligned} \tag{4.80}$$

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

Proof Using similar arguments as in Proposition 2.2 in [17], 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 KS 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 [80], 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 that

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

From Theorem 1.3, Supplement 1 in [9], there exists a Borel function $g : \mathbb{R} \to \mathbb{R}$ with

$$g(\lambda) = \begin{cases} 1, & \text{if } \lambda \leq \lambda_N, \\ 0, & \text{otherwise,} \end{cases} \tag{4.81}$$

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

$$\operatorname*{argmin}_{\gamma \in \mathcal{K}_N} E_{\text{band}}(u, \phi, \gamma) = g\big(H(\phi, u)\big). \tag{4.82}$$

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

the minimization over the subset

$$\mathcal{K}_N^{H(\phi,u)} = \big\{\gamma \in \mathcal{K}_N : \gamma = g\big(H(\phi,u)\big) \text{ for a Borel function } g \text{ over } \mathbb{R},\, 0 \le g \le 1\big\} \tag{4.83}$$

and observe that

$$\inf_{\gamma\in\mathcal{K}_N} E_{\text{band}}(u,\phi,\gamma) = \inf_{\gamma\in\mathcal{K}_N^{H(\phi,u)}} \mathrm{Tr}\big(H(\phi,u)\gamma\big). \qquad \blacksquare \tag{4.84}$$

It bears emphasis 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.

In the next two sections, we proceed to define the spectral discretization and the spatial discretization of the reformulated extended KS energy functional defined in (4.30).

4.4.2 Spatial discretization

We begin by discretizing problem (4.72) *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

$$\phi_j(\mathbf{r}) = \sum_{a=1}^{j} \phi_a e_a(\mathbf{r}). \tag{4.85}$$

The nuclear charge distribution is

$$b_j(\mathbf{r},\{\mathbf{R_1},\cdots,\mathbf{R_M}\}) = \sum_{a=1}^{j} b_a^{\{\mathbf{R_1},\cdots,\mathbf{R_M}\}} e_a(\mathbf{r}), \tag{4.86}$$

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

$$u_j(\mathbf{r}) = \sum_{a=1}^{j} u_a d_a(\mathbf{r}). \tag{4.87}$$

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

$$\gamma_j(\mathbf{r}_1, \mathbf{r}_2) = \sum_{a_1=1}^{j} \sum_{a_2=1}^{j} \gamma^j_{a_1,a_2} e_{a_1}(\mathbf{r}_1) e_{a_2}(\mathbf{r}_2), \tag{4.88}$$

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

$$\rho_j(\mathbf{r}) = \sum_{a_1=1}^{j} \sum_{a_2=1}^{j} \rho^j_{a_1 a_2} e_{a_1}(\mathbf{r}) e_{a_2}(\mathbf{r}), \tag{4.89}$$

where

$$\rho^j_{a_1 a_2} = \gamma^j_{a_1,a_2}. \tag{4.90}$$

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

$$\mathcal{K}^j_N = \big\{\gamma \in \mathcal{X} : \gamma \in \mathcal{S}(\mathcal{V}_j),\ 0 \leq \gamma \leq 1\big\}, \tag{4.91}$$

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 (4.31) to $\mathcal{U}_j \times \mathcal{V}_j \times \mathcal{K}^j_N$, follow as

$$L^j(u, \phi, \gamma) = \mathrm{Tr}(H^j(\phi, u)\gamma^j) + \sum_{a_1=1}^{j} \sum_{a_2=1}^{j} \Big\{ -C_S \phi_{a_1} A_{a_1,a_2} \phi_{a_2} + b^{\{\mathbf{R_1},\cdots,\mathbf{R_M}\}}_{a_1} \mathcal{M}_{a_1,a_2} \phi_{a_2} \Big\} + B^*_{\mathrm{xc}}(u). \tag{4.92}$$

Before proceeding further, we remark on the notation in (4.92). 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 chapter, 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 these discrete quantities:

$$\begin{aligned} H^j &\equiv \frac{1}{2}A + \Phi^j - U^j, \\ A_{a_1,a_2} &\equiv \int_\Omega \nabla e_{a_1}(\mathbf{r}) \cdot \nabla e_{a_2}(\mathbf{r})\,\mathrm{d}\mathbf{r}, \\ \mathcal{M}_{a_1,a_2} &\equiv \int_\Omega e_{a_1}(\mathbf{r}) \cdot e_{a_2}(\mathbf{r})\,\mathrm{d}\mathbf{r}, \\ \Phi^j_{a_1,a_2} &\equiv \int_\Omega \Big(\sum_{a=1}^{j} \phi_a e_a(\mathbf{r})\Big) e_{a_1}(\mathbf{r}) e_{a_2}(\mathbf{r})\,\mathrm{d}\mathbf{r}, \\ U^j_{a_1,a_2} &\equiv \int_\Omega \Big(\sum_{a=1}^{j} u_a d_a(\mathbf{r})\Big) e_{a_1}(\mathbf{r}) e_{a_2}(\mathbf{r})\,\mathrm{d}\mathbf{r}. \end{aligned} \tag{4.93}$$

Formally, A and $\mathcal{M}$ also depend on j, as they are restrictions of operators to $\{e_1, \ldots, e_j\}$. We omit this dependence here for simplicity of notation.

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

$$E^j_{\text{band}}(u, \phi, \gamma) = \mathrm{Tr}(H^j(\phi, u)\gamma^j). \tag{4.94}$$

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

$$\mathcal{K}_N^{H^j(\phi,u)} = \big\{\gamma \in \mathcal{K}_N^j \,:\, \gamma = g\big(H^j(\phi,u)\big) \text{ for a Borel function } g \text{ over } \mathbb{R},\, 0 \le g \le 1\big\}. \tag{4.95}$$

With these settings, motivated by the equations (4.30)–(4.33), the corresponding sequence

of discrete energies $\epsilon_{0,j}^{\mathrm{REKS}}$ becomes

$$\epsilon_{0,j}^{\mathrm{REKS}} = \inf_{u\in\mathcal{U}_j}\sup_{\phi\in\mathcal{V}_j}\inf_{\gamma\in\mathcal{K}_N^{H^j(\phi,u)}} L^j(u,\phi,\gamma). \tag{4.96}$$

4.4.3 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 (4.94). We begin by applying the spectral decomposition theorem (cf., e.g., [65]). For fixed $j \in \mathbb{N}$, since H^j defined in (4.93) is a self-adjoint operator, this theorem states that

$$H^j = \int_{\sigma(H^j)} \lambda \,\mathrm{d}P^j(\lambda), \tag{4.97}$$

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 (4.88) defined for H^j, there exist bounded Borel functions $g^j : \mathbb{R} \to \mathbb{R}$ with

$$\gamma^j = \int_{\sigma(H^j)} g^j(\lambda)\,\mathrm{d}P^j(\lambda). \tag{4.98}$$

Using this representation, we define

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

and where

$$\mu^j_{e_a,e_a}(\lambda) \equiv \langle e_a|P^j(\lambda)|e_a\rangle \tag{4.99}$$

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

$$\mu^j_{e_a,e_a}(\lambda) = \begin{cases} 0 & \text{if } \lambda < \lambda_1, \\ \langle e_a|P^j(\lambda_k)|e_a\rangle & \text{if } \lambda_k \leq \lambda < \lambda_{k+1}, k = 1, \ldots, j-1, \\ \langle e_a|P^j(\lambda_j)|e_a\rangle & \text{if } \lambda \geq \lambda_j. \end{cases} \tag{4.100}$$

Knowing the quantities $E^j_{\text{band}}(g^j)$, $N^j(g^j)$ 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

$$\inf_{g^j \in \mathfrak{B}} \left\{ E^j_{\text{band}}(g^j),\ 0 \leq g^j \leq 1,\ N^j(g^j) = N \right\}, \tag{4.101}$$

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 (4.101) by applying the Rayleigh-Ritz method over a finite-dimensional subspace $\mathfrak{B}_k$ of $\mathfrak{B}$ spanned by a chosen spectral basis $\{s^k_1, \ldots, s^k_k\}$, $k \in \mathbb{N}$. Any basis that spans the space of real-valued bounded measurable functions can be chosen for spectral discretization. In practice, it is advantageous to choose a basis in which its spectral integral for each $e_a, a \in \mathbb{N}$,

$$\int_{\sigma(H^j)} s^k_q(\lambda) d\mu^j_{e_a,e_a}(\lambda), \tag{4.102}$$

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

$$\mathcal{K}^{H^j(\phi,u)}_{N,k} = \left\{ \gamma \in \mathcal{K}^{H^j(\phi,u)}_N : \gamma = \sum_{q=1}^{k} c^k_q s^k_q(H^j) \right\}. \tag{4.103}$$

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

$$\begin{aligned} E^j_{\text{band}}(\gamma) &= E^j_{\text{band}}\Big(\sum_{q=1}^{k} c_q^k s_q^k\Big) = \text{Tr}(H^j\gamma) \\ &= \sum_{i=1}^{\infty} \int_{\sigma(H^j)} \lambda \sum_{q=1}^{k} c_q^k s_q^k(\lambda)\, \mathrm{d}\mu^j_{e_i,e_i}(\lambda) \\ &= \sum_{q=1}^{k} c_q^k \left\{ \sum_{i=1}^{\infty} \int_{\sigma(H^j)} \lambda s_q^k(\lambda)\, \mathrm{d}\mu^j_{e_i,e_i}(\lambda) \right\} \equiv \sum_{q=1}^{k} c_q^k w_q^{k,j}, \end{aligned} \tag{4.104}$$

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

$$\begin{aligned} N^j(\gamma) &= N^j\Big(\sum_{q=1}^{k} c_q^k s_q^k\Big) = \text{Tr}(\gamma) \\ &= \sum_{i=1}^{\infty} \int_{\sigma(H^j)} \sum_{q=1}^{k} c_q^k s_q^k(\lambda)\, \mathrm{d}\mu^j_{e_i,e_i}(\lambda) \\ &= \sum_{q=1}^{k} c_q^k \left\{ \sum_{i=1}^{\infty} \int_{\sigma(H^j)} s_q^k(\lambda)\, \mathrm{d}\mu^j_{e_i,e_i}(\lambda) \right\} \equiv \sum_{q=1}^{k} c_q^k n_q^{k,j}. \end{aligned} \tag{4.105}$$

The minimization of the energy function in equation (4.101) over $\mathfrak{B}_k$ becomes

$$\inf_{\{c_q^k\}\subset\mathbb{R}^k} E^j_{\text{band}}\Big(\sum_{q=1}^{k} c_q^k s_q^k\Big), \tag{4.106}$$

subject to the constraints

$$0 \le c_q^k \le 1, \quad \sum_{q=1}^{k} c_q^k n_q^{k,j} = N. \tag{4.107}$$

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

4.4.3.1 Spectral binning

Spectral binning refers to a basis consisting of a collection of disjoint piecewise constant functions, also known as *simple functions*. The spectral binning basis is defined over a partition of the fixed interval $[\lambda_{LB}, \lambda_{UB}]$ into k sub-intervals, or *bins*, $\{t_q^k, q = 0, \ldots, k\}$.

We require that $t_0^k = \lambda_{\mathrm{LB}} \leq \lambda_{\min}$ and $\lambda_N \leq \lambda_{\mathrm{UB}} = t_k^k < \lambda_{\max}$, where $\lambda_{\min}$ and $\lambda_{\max}$ are the minimum and maximum eigenvalues of H^j, respectively. The choice of $(\lambda_{\mathrm{LB}}, \lambda_{\mathrm{UB}})$ must ensure that the space $\mathcal{K}_{N,k}^{H^j(\phi,u)}$ includes the minimizer $\gamma_{\min}$ to the band energy functional $E_{\mathrm{band}}^j(g^j)$. Let $s_{t_q^k}(\lambda)$ denote the disjoint piecewise constant characteristic functions defined on the spectrum of $H^j(\phi, u)$,

$$s_{t_q^k}(\lambda) \equiv \begin{cases} 1, & \text{if } t_{q-1}^k \leq \lambda \leq t_q^k, \\ 0, & \text{otherwise.} \end{cases} \tag{4.108}$$

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 $[\lambda_{\mathrm{LB}}, \lambda_{\mathrm{UB}}]$. The density matrix $\gamma_k^j \in \mathcal{K}_{N,k}^{H^j(\phi,u)}$ using the spectral theorem in the spectral binning basis is

$$\gamma_k^j = \int_{\sigma(H^j)} \sum_{q=1}^k c_q^k s_{t_q^k}(\lambda) \, \mathrm{d}P^j(\lambda). \tag{4.109}$$

For any $\gamma \in \mathcal{K}_{N,k}^{H^j(\phi,u)}$ with associated coefficients $\{c_q^k\}_{q=1}^k$ as in equation (4.109), the corresponding band energy is

$$\begin{aligned} E_{\mathrm{band}}^j(\gamma) &= E_{\mathrm{band}}^j\Big(\sum_{q=1}^k c_q^k s_{t_q^k}\Big) = \mathrm{Tr}(H^j\gamma) \\ &= \sum_{q=1}^k c_q^k \left(\sum_{i=1}^\infty \int_{\sigma(H^j)} \lambda s_{t_q^k}(\lambda) \, \mathrm{d}\mu_{e_i,e_i}^j(\lambda) \right) = \sum_{q=1}^k c_q^k w_q^{k,j}, \end{aligned}$$

and

$$\begin{aligned} N^j(\gamma) &= N^j\Big(\sum_{q=1}^k c_q^k s_{t_q^k}\Big) = \mathrm{Tr}(\gamma) \\ &= \sum_{q=1}^k c_q^k \left(\sum_{i=1}^\infty \int_{\sigma(H^j)} s_{t_q^k}(\lambda) \, \mathrm{d}\mu_{e_i,e_i}^j(\lambda) \right) = \sum_{q=1}^k c_q^k n_q^{k,j}, \end{aligned}$$

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

The minimization over $\mathfrak{B}_k$ in equation (4.101) becomes a linear programming problem,

$$\inf_{\{c_q^k\}\subset\mathbb{R}^k} \sum_{q=1}^{k} c_q^k w_q^{k,j}, \tag{4.110}$$

subject to the linear constraints

$$0 \le c_q^k \le 1, \quad \sum_{q=1}^{k} c_q^k n_q^{k,j} = N. \tag{4.111}$$

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

4.4.3.2 Numerical evaluation of $\{n_q^{k,j}\}_{q=1}^k$

By Sylvester's law of inertia [75], $n_q^{k,j}$ equals the number of eigenvalues of $H^j(\phi, u)$ contained in the sub-interval (t_{q-1}^k, t_q^k). 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 that 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, [54]. To find the number of eigenvalues of the discrete Hamiltonian matrix H^j in an interval $[t_{q-1}^k, t_q^k]$, we need to perform the LDL^T decomposition twice:

$$\begin{aligned} H^j - t_{q-1}^k \mathcal{I}^j &= L_{t_{q-1}^k} D_{t_{q-1}^k} L_{t_{q-1}^k}^T, \\ H^j - t_q^k \mathcal{I}^j &= L_{t_q^k} D_{t_q^k} L_{t_q^k}^T. \end{aligned} \tag{4.112}$$

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 $\mathcal{N}_-(D_{t_q^k})$ denote the number of negative eigenvalues of $D_{t_q^k}$. Then,

$$n_q^k = \mathcal{N}_-(D_{t_q^k}) - \mathcal{N}_-(D_{t_{q-1}^k}). \tag{4.113}$$

Turning to the computational cost for the LDL^T decomposition, we note that for a $j \times j$ matrix with half bandwidth W, the number of operations for the LDL^T decomposition is [54],

$$C_{LDL^T} = \frac{W(W+1)j}{2}. \tag{4.114}$$

Thus, for k partitions or "bins" of the spectrum, the total number of operations to obtain the number of eigenvalues in each bin is

$$C_{\text{binning}} = \frac{W(W+1)kj}{2}. \tag{4.115}$$

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 [44], the computational cost for the LDL^T decomposition of a molecular system in 3D at worst scales as N^2. Note that by (4.115), the computational cost of the binning method scales linearly with respect to the number of spectral discretizations k.

4.4.4 Numerical evaluation of $\{w_q^{k,j}\}_{q=1}^k$

Unlike $n_q^{k,j}$ introduced in (4.105), it is not possible to evaluate $w_q^{k,j}$ defined in (4.104) directly at a cost that scales better than cubic with respect to the number of electrons in the system. Therefore, we proceed to make one more approximation. Let $\{m_q^k\}_{q=1}^k$ be the center of mass

of each partition, defined by

$$m_q^k \equiv \frac{w_q^{k,j}}{n_q^{k,j}} = \frac{1}{n_q^{k,j}} \sum_{i=1}^{\infty} \left(\int_{\sigma(H^j)} \lambda s_{t_q^k}(\lambda) \, \mathrm{d}\mu_{e_i,e_i}^j(\lambda) \right). \tag{4.116}$$

We approximate the center of mass m_q^k in the interval (t_{q-1}^k, t_q^k) by

$$m_q^k \approx \frac{t_q^k - t_{q-1}^k}{2}. \tag{4.117}$$

This approximation implies the spectral approximation of the band energy as

$$\begin{aligned} \mathrm{Tr}\big(H^j(\phi,u)\gamma^j\big) &= \sum_{i=1}^{\infty} \int_{\sigma(H^j)} \sum_{q=1}^{k} c_q \lambda s_{t_q^k}(\lambda) \, \mathrm{d}\mu_{e_i,e_i}^j(\lambda) \\ &\approx \sum_{q=1}^{k} c_q m_q^{k,j} n_q^{k,j} \equiv \tilde{\mathrm{Tr}}(H^j(\phi,u)\gamma^j). \end{aligned} \tag{4.118}$$

This approximation of $\{w_q^{k,j}\}_{q=1}^k$ introduces an error over the Rayleigh-Ritz approximation of the discrete band energy. However, in the following section we show that this error is controllable.

4.5 Convergence with respect to spectral and spatial discretization

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

Part I: Definition of the limit functionals.

Starting from equation (4.80), we consider the minimization problem

$$\epsilon_0^{\mathrm{R}EKS} = \inf_{u \in \mathcal{U}} T(u), \tag{4.119}$$

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

$$T(u) = B^*_{\text{xc}}(u) + \sup_{\phi \in \mathcal{V}} S(u, \phi), \tag{4.120}$$

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

$$S(u,\phi) = -\int_\Omega \left(C_S|\nabla\phi(\mathbf{r})|^2 - b(\mathbf{r}, \{\mathbf{R_1}, \cdots, \mathbf{R_M}\})\phi(\mathbf{r})\right) \mathrm{d}\mathbf{r} + \inf_{\gamma \in \mathcal{X}} \left\{ E_{\text{band}}(u,\phi,\gamma) + I_{\mathcal{K}_N^{H(\phi,u)}}(\gamma) \right\}. \tag{4.121}$$

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

$$I_{\mathcal{M}}(u) \equiv \begin{cases} 0 & \text{if } u \in \mathcal{M}, \\ +\infty & \text{otherwise.} \end{cases} \tag{4.122}$$

In (4.121), the minimization over $\mathcal{K}_N$ is replaced by the minimization over $\mathcal{K}_N^{H(\phi,u)}$. This ensures the existence of a spectral function and is justified in equation (4.84).

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

For $j \in \mathbb{N}$, based on the identity (4.80), we introduce the family of energies

$$\epsilon_{j,k_j} = \inf_{u \in \mathcal{U}} T^{j,k_j}(u), \tag{4.123}$$

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

$$T^{j,k_j}(u) = B^*_{\text{xc}}(u) + \sup_{\phi \in \mathcal{V}} S^{j,k_j}(u,\phi) + I_{\mathcal{U}_j}(u), \tag{4.124}$$

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

$$S^{j,k_j}(u,\phi) = -\int_\Omega \left(C_S|\nabla\phi(\mathbf{r})|^2 - b(\mathbf{r}, \{\mathbf{R_1}, \cdots, \mathbf{R_M}\})\phi(r)\right) \mathrm{d}\mathbf{r} + \inf_{\gamma \in \mathcal{X}} \left\{ E_{\text{band}_{j,k_j}}(u,\phi,\gamma) + I_{\mathcal{K}_{N,k_j}^{Hj(\phi,u)}}(\gamma) \right\} - I_{\mathcal{V}_j}(\phi). \tag{4.125}$$

In (4.125), we introduced the approximated constrained sets of density matrices

$$\mathcal{K}_{N,k_j}^{H^j(\phi,u)} = \Big\{\gamma \in \mathcal{K}_N : \gamma = \sum_{i=1}^{k_j} c_i^{k_j} s_{t_i^{k_j}}(H^j),\ 0 \le c_i^{k_j} \le 1\Big\} \tag{4.126}$$

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

$$E_{\mathrm{band}_{j,k_j}}(u,\phi,\gamma) = \tilde{\mathrm{Tr}}\big(H^j(\phi,u)\gamma\big), \tag{4.127}$$

where $\tilde{\mathrm{Tr}}(\cdot)$ (depending on k_j) is the approximation of the trace operator described in equation (4.118). We emphasize that this is the actual numerical approximation of the binning algorithm introduced in Section 4.4.4.

Summarizing (4.104) and (4.118), for $\gamma_{k_j} \in \mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}$ the approximate trace operator is

$$\begin{aligned}\tilde{\mathrm{Tr}}(H^j\gamma_{k_j}) &= \sum_{i=1}^{\infty}\sum_{q=1}^{k_j} c_q^{k_j} m_q^{k_j} \int_{t_q^{k_j}}^{t_{q+1}^{k_j}} s_q^{k_j}(\lambda)\,\mathrm{d}\mu_{e_i,e_i}(\lambda) \\ &= \sum_{i=1}^{\infty}\sum_{q=1}^{k_j} c_q^{k_j} m_q^{k_j}\big(\mu_{e_i,e_i}(t_{q+1}^{k_j}) - \mu_{e_i,e_i}(t_q^{k_j})\big),\end{aligned} \tag{4.128}$$

where $m_q^{k_j} \equiv \frac{t_{q+1}^{k_j}+t_q^{k_j}}{2}$ denotes as in (4.117) 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 $\mathrm{Tr}(H^j(\phi_j,u_j)\gamma_j)$. Then, in Section 4.5.2, we validate the convergence of the approximate trace operators.

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

Lemma 4.5.1 *If $u_j \rightharpoonup u$ in $\mathcal{U}$ and $\phi_j \rightharpoonup \phi$ in $\mathcal{V}$, then*

$$\liminf_{j\to\infty}\Big\{\mathrm{Tr}\big(H^j(u_j,\phi_j)\gamma_j\big) + I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma)\Big\} \ge E_{\mathrm{band}}(u,\phi,\gamma) + I_{\mathcal{K}_N^{H(\phi,u)}}(\gamma) \tag{4.129}$$

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 4.3.6,

$$\liminf_{j\to\infty} \mathrm{Tr}(-\Delta\gamma_j) \geq \mathrm{Tr}(-\Delta\gamma), \tag{4.130}$$

   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

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

   leading to

$$\liminf_{j\to\infty} \mathrm{Tr}\big(H^j(\phi_j,u_j)\gamma_j\big) \geq \mathrm{Tr}\big(H(\phi,u)\gamma\big). \tag{4.131}$$

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

$$\liminf_{j\to\infty} \Big\{\mathrm{Tr}\big(H^j(u_j,\phi_j)\gamma_j\big) + I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma)\Big\} = +\infty \geq E_{\mathrm{band}}(u,\phi,\gamma) + I_{\mathcal{K}_N^{H(\phi,u)}}(\gamma). \tag{4.132}$$

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

$$\liminf_{j\to\infty}\left\{\mathrm{Tr}\big(H^j(u_j,\phi_j)\gamma_j\big)+I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma)\right\}=E_{\mathrm{band}}(u,\phi,\gamma)+I_{\mathcal{K}_N^{H(\phi,u)}}(\gamma)=+\infty. \tag{4.133}$$

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 $i\in\mathbb{N}$,

$$\lim_{j\to\infty}\|\xi_i^j-\xi_i\|_{\mathcal{L}^2(\Omega)}=0,\qquad\lim_{j\to\infty}\xi_i^j=\xi_i. \tag{4.134}$$

   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. Therefore,

$$\gamma\xi_1=\sum_{q=1}^{\infty}c_{1q}\xi_q, \tag{4.135}$$

   and there must exist an index $p\in\mathbb{N}$, $p\neq1$, such that $c_{1p}\neq0$. Consider this c_{1p}. Then

$$c_{1p}=\langle\gamma\xi_1,\xi_p\rangle=\lim_{j\to\infty}\langle\gamma_j\xi_1,\xi_p\rangle=\lim_{j\to\infty}\langle g_j(H^j)\xi_1,\xi_p\rangle. \tag{4.136}$$

   Therefore, for $p\neq1$,

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

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 $\mathcal{X}$,

$$\liminf_{j\to\infty} \left\{ \mathrm{Tr}\big(H^j(u_j,\phi_j)\gamma_j\big) + I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma) \right\} \geq E_{\mathrm{band}}(u,\phi,\gamma) + I_{\mathcal{K}_N^{H(\phi,u)}}(\gamma). \qquad \blacksquare \quad (4.137)$$

Lemma 4.5.2 *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*

$$\limsup_{j\to\infty} \mathrm{Tr}\big(H^j(u_j,\phi_j)\gamma_j\big) \leq E_{\mathrm{band}}(u,\phi,\gamma) \quad (4.138)$$

and

$$\mathrm{Tr}\big(H^j(u_j,\phi_j)\gamma\big) + I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma) \overset{\Gamma}{\rightharpoonup} E_{\mathrm{band}}(u,\phi,\gamma) + I_{\mathcal{K}_N^{H(\phi,u)}}(\gamma) \quad (4.139)$$

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, we trivially have
$$\limsup_{j\to\infty} \mathrm{Tr}\big(H^j(u_j,\phi_j)\gamma_j\big) \leq E_{\mathrm{band}}(u,\phi,\gamma) = +\infty. \quad (4.140)$$

2. If $\gamma \in \mathcal{K}_N^{H(\phi,u)}$, then without loss of generality, we write
$$\gamma = \sum_{i=1}^{\infty} 2\alpha_i \xi_i\rangle\langle\xi_i, \quad (4.141)$$

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 4.5.1.

Let us define the sequence of finite-rank operators:

$$\gamma_j = \sum_{i=1}^{j} 2\alpha_i \xi_i^j \rangle\langle \xi_i^j. \tag{4.142}$$

We proceed to show that $\gamma_j \to \gamma$ w.r.t. $\|\cdot\|_{\mathcal{X}}$. From Theorem VI.10 in [63], there exists an unique partial isometry Q, such that

$$|\gamma - \gamma_j| = Q(\gamma - \gamma_j). \tag{4.143}$$

Now we show the strong convergence of $\gamma_j \to \gamma$ in the norm sense of $\mathcal{X}$ as follows. Utilizing equation (4.143), the dual operator Q^* of Q, the Cauchy-Schwarz inequality and the fact that both Q and Q^* are isometries, we find

$$\begin{aligned}
\lim_{j\to\infty} \mathrm{Tr}(|\gamma-\gamma_j|) &= \lim_{j\to\infty} \mathrm{Tr}(Q(\gamma-\gamma_j)) \\
&= \lim_{j\to\infty} \sum_{p=1}^{\infty} \langle Q(\gamma-\gamma_j)\xi_p, \xi_p\rangle \\
&= \lim_{j\to\infty} \sum_{p=1}^{\infty} \langle (\gamma-\gamma_j)\xi_p, Q^*\xi_p\rangle \\
&\leq \lim_{j\to\infty} \sum_{p=1}^{\infty} \|(\gamma-\gamma_j)\xi_p\|_{\mathcal{L}^2(\Omega)} \|Q^*\xi_p\|_{\mathcal{L}^2(\Omega)} \\
&\leq \lim_{j\to\infty} \sum_{p=1}^{\infty} \|(\gamma-\gamma_j)\xi_p\|_{\mathcal{L}^2(\Omega)} \|\xi_p\|_{\mathcal{L}^2(\Omega)} \\
&= \lim_{j\to\infty} \sum_{p=1}^{\infty} \|(\gamma-\gamma_j)\xi_p\|_{\mathcal{L}^2(\Omega)}.
\end{aligned} \tag{4.144}$$

Let us consider just one of the terms in equation (4.144) for fixed summation index p. We now look at its projection onto the eigen-basis $\{\xi_i\}_{i\in\mathbb{N}}$ and find with (4.141),

(4.142)

$$\begin{aligned}
\lim_{j\to\infty} \|(\gamma-\gamma_j)\xi_p\|^2_{\mathcal{L}^2(\Omega)} &= \lim_{j\to\infty} \sum_{q=1}^{\infty} \Big|\langle(\gamma-\gamma_j)\xi_p,\xi_q\rangle\Big|^2 \\
&= \lim_{j\to\infty} \left\{ \sum_{q=1}^{\infty} \left| 2\alpha_q\langle\xi_p,\xi_q\rangle - \sum_{i=1}^{j} 2\alpha_i\langle\xi_p,\xi_i^j\rangle\langle\xi_i^j,\xi_q\rangle \right|^2 \right\} \\
&\le \lim_{j\to\infty} \left\{ \left| 2\alpha_p - \sum_{i=1}^{j} 2\alpha_i\langle\xi_p,\xi_i^j\rangle^2 \right|^2 + \sum_{q=1,q\neq p}^{\infty} \left| \sum_{i=1}^{j} 2\alpha_i\langle\xi_p,\xi_i^j\rangle\langle\xi_i^j,\xi_q\rangle \right|^2 \right\} \\
&\le \lim_{j\to\infty} \left\{ \left| 2\alpha_p - \sum_{i=1}^{j} 2\alpha_i\langle\xi_p,(\xi_i^j-\xi_i)+\xi_i\rangle^2 \right|^2 \right. \\
&\qquad \left. + \sum_{q=1,q\neq p}^{\infty} \left| \sum_{i=1}^{j} 2\alpha_i\langle\xi_p,(\xi_i^j-\xi_i)+\xi_i\rangle\langle(\xi_i^j-\xi_i)+\xi_i,\xi_q\rangle \right|^2 \right\} = 0.
\end{aligned} \tag{4.145}$$

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

$$\lim_{j\to\infty} \langle\xi_i,\xi_q^j-\xi_q\rangle = \lim_{j\to\infty} \langle\xi_q^j-\xi_q,\xi_i\rangle = 0. \tag{4.146}$$

With the help of (4.145), we find that

$$0 = \sum_{p=1}^{\infty} \liminf_{j\to\infty} \|(\gamma-\gamma_j)\xi_p\|_{\mathcal{L}^2(\Omega)} \le \liminf_{j\to\infty} \sum_{p=1}^{\infty} \|(\gamma-\gamma_j)\xi_p\|_{\mathcal{L}^2(\Omega)}. \tag{4.147}$$

Similarly, by Jensen's inequality,

$$\limsup_{j\to\infty} \sum_{p=1}^{\infty} \|(\gamma-\gamma_j)\xi_p\|_{\mathcal{L}^2(\Omega)} \le \sum_{p=1}^{\infty} \limsup_{j\to\infty} \|(\gamma-\gamma_j)\xi_p\|_{\mathcal{L}^2(\Omega)} = 0. \tag{4.148}$$

As a result of (4.147), (4.148) we have $0 \le \liminf_{j\to\infty} \|(\gamma-\gamma_j)\xi_p\|_{\mathcal{L}^2(\Omega)} \le \limsup_{j\to\infty} \|(\gamma-\gamma_j)\xi_p\|_{\mathcal{L}^2(\Omega)} \le 0$, implying that

$$\lim_{j\to\infty} \mathrm{Tr}(|\gamma-\gamma_j|) \le \lim_{j\to\infty} \|(\gamma-\gamma_j)\xi_p\|_{\mathcal{L}^2(\Omega)} = 0. \tag{4.149}$$

We proceed to approximate each γ_j using spectral theory. By the choice of γ_j, there are suitable bounded Borel functions g_j such that

$$\gamma_j = g_j(H^j). \tag{4.150}$$

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

$$\tilde{\gamma}_{j,k} = \sum_{i=1}^{k} c_i^{k,j}\, s_{t_i^k}(H^j), \tag{4.151}$$

where

$$c_i^{k,j} \equiv \max\{g_j(t_i^k), g_j(t_{i+1}^k)\}, \tag{4.152}$$

and $\{t_1^k, \ldots, t_k^k\}$ is the partition of the interval $[\lambda_{\mathrm{LB}}, \lambda_{\mathrm{UB}}]$ introduced in Section 4.4.3.1. We can show that for every $j \in \mathbb{N}$

$$\mathrm{Tr}(|\tilde{\gamma}_{j,k} - \gamma_j|) \to 0 \tag{4.153}$$

as $k \to \infty$, see Theorem 2.29 in [82]. However, the trace of $\tilde{\gamma}_{j,k}$ does not satisfy the trace condition for every k, i.e.,

$$\mathrm{Tr}(\tilde{\gamma}_{j,k}) \neq N. \tag{4.154}$$

Nevertheless, since

$$\lim_{k\to\infty} \mathrm{Tr}(\tilde{\gamma}_{j,k}) = N, \tag{4.155}$$

we can normalize the trace to N by introducing

$$\gamma_{j,k} \equiv \frac{N}{\mathrm{Tr}(\tilde{\gamma}_{j,k})}\tilde{\gamma}_{j,k}, \tag{4.156}$$

Here, due to (4.153), we may assume $\mathrm{Tr}(\tilde{\gamma}_{j,k}) \neq 0$ for all j and k.

In conclusion, we have

$$\lim_{k\to\infty} \mathrm{Tr}(|\gamma_{j,k} - \gamma_j|) \leq \lim_{k\to\infty} \big\{\mathrm{Tr}(|\gamma_{j,k} - \tilde{\gamma}_{j,k}|) + \mathrm{Tr}(|\tilde{\gamma}_{j,k} - \gamma_j|)\big\} = 0. \tag{4.157}$$

Eqn. (4.157) implies that for every j there is an index $k_j \in \mathbb{N}$, $k_j \to \infty$ as $j \to \infty$, such that

$$\mathrm{Tr}(|\gamma_{k_j} - \gamma_j|) \leq \frac{1}{j}. \tag{4.158}$$

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

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

Now, in order to show that

$$\mathrm{Tr}\big(\big||\nabla|(\gamma_{k_j} - \gamma)|\nabla|\big|\big) \to 0 \tag{4.159}$$

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

$$\lim_{j\to\infty} \|\gamma_{k_j} - \gamma\|_{\sup} \leq \lim_{j\to\infty} \mathrm{Tr}(|\gamma_{k_j} - \gamma|) = 0. \tag{4.160}$$

Combining the above arguments, it follows that

$$\begin{aligned}\liminf_{j\to\infty} \mathrm{Tr}\big(\big||\nabla|(\gamma_{k_j} - \gamma)|\nabla|\big|\big) &= \liminf_{j\to\infty} \mathrm{Tr}(Q|\nabla|(\gamma_{k_j} - \gamma)|\nabla|) \\ &= \liminf_{j\to\infty} \sum_{q=1}^{\infty} \langle Q|\nabla|(\gamma_{k_j} - \gamma)|\nabla|\xi_q, \xi_q\rangle \\ &\geq \sum_{q=1}^{\infty} \liminf_{j\to\infty} \langle(\gamma_{k_j} - \gamma)|\nabla|\xi_q, |\nabla|Q^*\xi_q\rangle = 0,\end{aligned} \tag{4.161}$$

and similarly

$$\begin{aligned}\limsup_{j\to\infty}\mathrm{Tr}\big(\big||\nabla|(\gamma_{k_j}-\gamma)|\nabla|\big|\big) &= \limsup_{j\to\infty}\mathrm{Tr}(Q|\nabla|(\gamma_{k_j}-\gamma)|\nabla|)\\ &= \limsup_{j\to\infty}\sum_{q=1}^{\infty}\langle Q|\nabla|(\gamma_{k_j}-\gamma)|\nabla|\xi_q,\xi_q\rangle\\ &\leq \sum_{q=1}^{\infty}\limsup_{j\to\infty}\langle(\gamma_{k_j}-\gamma)|\nabla|\xi_q,|\nabla|Q^*\xi_q\rangle = 0. \qquad (4.162)\end{aligned}$$

Together, (4.161) and (4.162) yield

$$\lim_{j\to\infty}\mathrm{Tr}\big(\big||\nabla|(\gamma_{k_j}-\gamma)|\nabla|\big|\big) = 0. \qquad (4.163)$$

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:

$$\limsup_{j\to\infty}\mathrm{Tr}\big(H^j(u_j,\phi_j)\gamma_j\big) = E_{\mathrm{band}}(u,\phi,\gamma), \qquad (4.164)$$

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 $j\to\infty$,

$$\mathrm{Tr}\big(H^j(u_j,\phi_j)\gamma\big)+I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j}}(\gamma)\overset{\Gamma}{\rightharpoonup}E_{\mathrm{band}}(u,\phi,\gamma)+I_{\mathcal{K}_N^{H(\phi,u)}}(\gamma). \qquad (4.165)$$ ∎

Lemma 4.5.3 *For every $\phi_j\rightharpoonup\phi$ in $\mathcal{V}$ and every $u_j\rightharpoonup u$ in $\mathcal{U}$, the family of functionals*

$$\Big\{\mathrm{Tr}\big(H^j(u_j,\phi_j)\gamma\big)+I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma)\Big\}_{j\in\mathbb{N}} \qquad (4.166)$$

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

Proof This proof is similar to the proof of Lemma 4.4.1. 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:

$$\begin{aligned}
\mathrm{Tr}\big(H^j(u_j,\phi_j)\gamma\big) &= \frac{1}{2}\mathrm{Tr}(-\Delta\gamma) + \mathrm{Tr}(\Phi_j\gamma) - \mathrm{Tr}(U_j\gamma) \\
&\geq \frac{1}{2}\mathrm{Tr}(-\Delta\gamma) - (\|\phi_j\|_{\mathcal{L}^2(\Omega)} + \|u_j\|_{\mathcal{U}})\|\rho_\gamma\|_{\mathcal{L}^2(\Omega)} \\
&\geq \frac{1}{2}\mathrm{Tr}(-\Delta\gamma) - C_{10}(\|\phi\|_{\mathcal{L}^2(\Omega)} + \|u_j\|_{\mathcal{L}^2(\Omega)})\|\rho_\gamma\|_{\mathcal{L}^1(\Omega)}^{\frac{1}{4}}\|\rho_\gamma\|_{\mathcal{L}^3(\Omega)}^{\frac{3}{4}} \qquad (4.167) \\
&\geq \frac{1}{2}\mathrm{Tr}(-\Delta\gamma) - C_{11}(\|\phi_j\|_{\mathcal{L}^2(\Omega)} + \|u_j\|_{\mathcal{L}^2(\Omega)})N^{1/4}\|\nabla\sqrt{\rho_\gamma}\|_{\mathcal{L}^2(\Omega)}^{\frac{3}{2}} \qquad (4.168) \\
&\geq \frac{1}{2}\mathrm{Tr}(-\Delta\gamma) - C_{12}\|\nabla\sqrt{\rho_\gamma}\|_{\mathcal{L}^2(\Omega)}^{\frac{3}{2}}, \qquad (4.169)
\end{aligned}$$

where interpolation inequalities are used to obtain (4.167), and the Gagliardo–Nirenberg–Sobolev inequality is used to obtain (4.168), and with the constant

$$C_{12} \equiv C_{11}\sup_{j\in\mathbb{N}}\Big\{\|\phi_j\|_{\mathcal{L}^2(\Omega)} + \|u_j\|_{\mathcal{L}^2(\Omega)}\Big\}N^{1/4}. \qquad (4.170)$$

Since

$$\mathrm{Tr}(-\Delta\gamma) \geq \|\nabla\sqrt{\rho_\gamma}\|_{\mathcal{L}^2(\Omega)}^2, \qquad (4.171)$$

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

$$\Big\{\gamma \in \mathcal{X} \,:\, \mathrm{Tr}\big(H^j(u_j,\phi_j)\gamma\big) + I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma) \leq t\Big\} \qquad (4.172)$$

are bounded:

$$t \geq \frac{1}{2}\|\gamma\|_{\mathcal{X}} - C_{12}\|\sqrt{\rho_\gamma}\|_{\mathcal{L}^2(\Omega)}^{\frac{3}{2}} - \frac{N}{2}. \qquad (4.173)$$

By the results in [36], 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 4.5.4 *If $\phi_j \rightharpoonup \phi$ in $\mathcal{V}$ and $u_j \rightharpoonup u$ in $\mathcal{U}$, then*

$$\lim_{j\to\infty} \inf_{\gamma\in\mathcal{X}} \left\{ \mathrm{Tr}\big(H^j(u_j,\phi_j)\gamma\big) + I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma) \right\} = \inf_{\gamma\in\mathcal{X}} \left\{ E_{\mathrm{band}}(u,\phi,\gamma) + I_{\mathcal{K}_N^{H(\phi,u)}}(\gamma) \right\}. \quad (4.174)$$

Proof This is proven using Theorem 7.8 in [49], Lemma 4.5.2, and Lemma 4.5.3. ∎

4.5.2 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 (4.127), i.e. to the evaluation operators actually used in the binning algorithm.

Lemma 4.5.5 *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*

$$\lim_{j\to\infty} \left| \tilde{\mathrm{Tr}}(H^j\gamma_{k_j}) - \mathrm{Tr}(H^j\gamma_{k_j}) \right| = 0. \quad (4.175)$$

Proof By direct estimates we find that

$$\begin{aligned}
\left| \tilde{\mathrm{Tr}}(H^j\gamma_{k_j}) - \mathrm{Tr}(H^j\gamma_{k_j}) \right| &= \left| \sum_{i=1}^{\infty}\sum_{q=1}^{k_j} \int_{t_q^{k_j}}^{t_{q+1}^{k_j}} c_q^{k_j}(m_q^{k_j}-\lambda) s_q^{k_j}(\lambda)\,\mathrm{d}\mu_{e_i,e_i}(\lambda) \right| \\
&= \left| \sum_{i=1}^{\infty}\sum_{q=1}^{k_j} c_q^{k_j}(m_q^{k_j}-\nu_{q,i}^{k_j}) \int_{t_q^{k_j}}^{t_{q+1}^{k_j}} s_q^{k_j}(\lambda)\,\mathrm{d}\mu_{e_i,e_i}(\lambda) \right| \quad (4.176) \\
&= \left| \sum_{i=1}^{\infty}\sum_{q=1}^{k_j} c_q^{k_j}(m_q^{k_j}-\nu_{q,i}^{k_j}) \big(\mu_{e_i,e_i}(t_{q+1}^{k_j}) - \mu_{e_i,e_i}(t_q^{k_j})\big) \right| \\
&\le \left| \sum_{q=1}^{k_j} c_q^{k_j} \frac{h_{k_j}}{2} \sum_{i=1}^{\infty} \big(\mu_{e_i,e_i}(t_{q+1}^{k_j}) - \mu_{e_i,e_i}(t_q^{k_j})\big) \right| \\
&= \left| \sum_{q=1}^{k_j} c_q^{k_j} \frac{h_{k_j}}{2} n_q^{k_j} \right|, \quad (4.177)
\end{aligned}$$

where $h_{k_j} := \max_{1\leq l\leq t_j-1} |t_l^{k_j} - t_{l+1}^{k_j}|$ 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 (4.176) 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., [82].

For each $\epsilon > 0$, there exists a $\overline{k} \in \mathbb{N}$ such that $h_{k_j} < \frac{2\epsilon}{N}$ for all $k_j \geq \overline{k}$. Consequently, due to equation (4.177),

$$\left|\tilde{\mathrm{Tr}}(H^j\gamma_{k_j}) - \mathrm{Tr}(H^j\gamma_{k_j})\right| < \left|\frac{\epsilon}{N}\sum_{q=1}^{k_j} c_q^{k_j} n_q^{k_j}\right| < \epsilon. \tag{4.178}$$

This concludes the proof of (4.175). ∎

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

$$\tilde{\mathrm{Tr}}\big(H^j(\phi_j,u_j)\gamma\big) + I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma) \xrightarrow{\Gamma} \mathrm{Tr}\big(H(\phi,u)\gamma\big) + I_{\mathcal{K}_N^{H(\phi,u)}}(\gamma) \tag{4.179}$$

in the limit $j \to \infty$.

Proof Let us begin with the liminf part of the Γ-convergence proof. From Lemma 4.5.1, 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 $\gamma_j \xrightarrow{*} \gamma$,

$$\mathrm{Tr}\big(H(\phi,u)\gamma\big) + I_{\mathcal{K}_N^{H(\phi,u)}}(\gamma) \leq \liminf_{j\to\infty} \Big\{\mathrm{Tr}\big(H^j(\phi_j,u_j)\gamma_j\big) + I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma_j)\Big\}. \tag{4.180}$$

Using Lemma 4.5.5,

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

Similarly, for the limsup part, using the same recovery sequence $\{\gamma_{k_j}\}_{j\in\mathbb{N}}$ as the one constructed in Lemma 4.5.2,

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

Therefore, using the results of Lemma 4.5.2,

$$\limsup_{j\to\infty}\Big\{\tilde{\mathrm{Tr}}\big(H^j(\phi_j,u_j)\gamma_{k_j}\big)+I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma_{k_j})\Big\}\le\mathrm{Tr}\big(H(\phi,u)\gamma\big)+I_{\mathcal{K}_N^{H(\phi,u)}}(\gamma).\qquad(4.181)$$

This completes the proof. ∎

Lemma 4.5.7 *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 4.5.5, we have for every $\gamma \in \mathcal{K}_{N,j,k_j}^{H^j(\phi_j,u_j)}$,

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

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

Hence, from Lemma 4.5.3, especially equation (4.169),

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

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

$$\left\{\gamma \in \mathcal{X} : \tilde{\mathrm{Tr}}\big(H^j(\phi_j,u_j)\gamma\big) + I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma) < t\right\} \tag{4.182}$$

are bounded:

$$t \geq \frac{1}{2}\|\gamma\|_{\mathcal{X}} - C_{12}\|\sqrt{\rho_\gamma}\|_{\mathcal{L}^2(\Omega)}^{\frac{3}{2}} - \frac{N}{2} + (\lambda_{\mathrm{LB}}-\lambda_{\mathrm{UB}})N. \qquad \blacksquare \tag{4.183}$$

Lemma 4.5.8 *If $\phi_j \rightharpoonup \phi$ in $\mathcal{V}$ and $u_j \rightharpoonup u$ in $\mathcal{U}$, then*

$$\lim_{j\to\infty}\inf_{\gamma\in\mathcal{X}}\left\{\tilde{\mathrm{Tr}}(H^j(\phi_j,u_j)\gamma) + I_{\mathcal{K}_{N,k_j}^{H^j(\phi_j,u_j)}}(\gamma)\right\} = \inf_{\gamma\in\mathcal{X}}\left\{\mathrm{Tr}\big(H(\phi,u)\gamma\big) + I_{\mathcal{K}_N^{H(\phi,u)}}(\gamma)\right\}. \tag{4.184}$$

Proof This is a direct consequence of Theorem 7.8 in [49], Lemma 4.5.6, and Lemma 4.5.7. ∎

4.5.3 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 4.5.9 *If $u_j \rightharpoonup u$ in $\mathcal{U}$, then for $j\to\infty$,*

$$-S^{j,k_j}(u_j,\phi) \xrightarrow{\Gamma} -S(u,\phi) \tag{4.185}$$

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

Proof From Lemma 4.5.8, for every $u\in\mathcal{U}$ and all $u_j \rightharpoonup u$ in $\mathcal{U}$,

$$\lim_{j\to\infty}\inf_{\gamma\in\mathcal{X}}\Big\{E_{\mathrm{band}_{j,k_j}}(u_j,\phi,\gamma)+I_{\mathcal{K}_{N,k_j}^{H^j(\phi,u_j)}}(\gamma)\Big\}=\inf_{\gamma\in\mathcal{X}}\Big\{E_{\mathrm{band}}(u,\phi,\gamma)+I_{\mathcal{K}_N^{H(\phi,u)}}(\gamma)\Big\}. \tag{4.186}$$

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

$$\int_\Omega C_S|\nabla\phi(\mathbf{r})|^2\,\mathrm{d}\mathbf{r}\le\liminf_{j\to\infty}\int_\Omega C_S|\nabla\phi_j(\mathbf{r})|^2\,\mathrm{d}\mathbf{r}, \tag{4.187}$$

and

$$-\int_\Omega b(\mathbf{r},\{\mathbf{R_1},\cdots,\mathbf{R_M}\})\phi(\mathbf{r})\,\mathrm{d}\mathbf{r}\le\liminf_{j\to\infty}\left(-\int_\Omega b(\mathbf{r},\{\mathbf{R_1},\cdots,\mathbf{R_M}\})\phi_j(\mathbf{r})\,\mathrm{d}\mathbf{r}\right). \tag{4.188}$$

This shows

$$-S(u,\phi)\le\liminf_{j\to\infty}\big(-S^{j,k_j}(u_j,\phi)\big). \tag{4.189}$$

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

$$\lim_{j\to\infty}\int_\Omega C_S|\nabla\tilde{\phi}(\mathbf{r})|^2\,\mathrm{d}\mathbf{r}=\int_\Omega C_S|\nabla\phi(\mathbf{r})|^2\,\mathrm{d}\mathbf{r} \tag{4.190}$$

and

$$\lim_{j\to\infty}\left(-\int_\Omega b(\mathbf{r},\{\mathbf{R_1},\cdots,\mathbf{R_M}\})\tilde{\phi}_j(\mathbf{r})\,\mathrm{d}\mathbf{r}\right)=-\int_\Omega b(\mathbf{r},\{\mathbf{R_1},\cdots,\mathbf{R_M}\})\phi(\mathbf{r})\,\mathrm{d}\mathbf{r}. \tag{4.191}$$

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 4.5.10 *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 4.5.3, we find

$$\begin{aligned}-S^{j,k_j}(u_j,\phi) \;=&\int_\Omega\big(C_S|\nabla\phi(\mathbf{r})|^2-b(\mathbf{r},\{\mathbf{R_1},\cdots,\mathbf{R_M}\})\phi(\mathbf{r})\big)\,\mathrm{d}\mathbf{r}\\ &-\inf_{\gamma\in\mathcal{X}}\Big\{\tilde{\mathrm{Tr}}\big(H^j(\phi,u_j)\gamma\big)+I_{\mathcal{K}_{N,k_j}^{H^j(\phi,u_j)}}(\gamma)\Big\}+I_{\mathcal{V}_j}(\phi) \qquad (4.192)\\ \geq\, &C_S\|\nabla\phi\|^2_{\mathcal{L}^2(\Omega)}-\|b(\mathbf{r},\{\mathbf{R_1},\cdots,\mathbf{R_M}\})\|_{\mathcal{L}^2(\Omega)}\|\phi\|_{\mathcal{L}^2(\Omega)}-\mathrm{Tr}\big(H^j(\phi,u_j)\hat{\gamma}_j\big)+\epsilon_{k_j}.\end{aligned}$$

Here, $\hat{\gamma}_j\in\mathcal{K}_{N,k_j}^{H^j(\phi,u_j)}$ are minimal in (4.192) and satisfy for all $j\in\mathbb{N}$

$$\tilde{\mathrm{Tr}}\big(H^j(\phi,u_j)\hat{\gamma}_j\big)=\mathrm{Tr}\big(H^j(\phi,u_j)\hat{\gamma}_j\big)-\epsilon_{k_j}, \tag{4.193}$$

where due to Lemma 4.5.5 the sequence ϵ_{k_j} converges to 0 as j becomes infinite. It follows

that

$$\begin{aligned} -S^{j,k_j}(u_j,\phi) \geq & C_{13}\|\phi\|^2_{\mathcal{L}^2(\Omega)} - (\|b(\mathbf{r},\{\mathbf{R_1},\cdots,\mathbf{R_M}\})\|_{\mathcal{L}^2(\Omega)} + \|\rho_{\hat{\gamma}_j}\|_{\mathcal{L}^2(\Omega)})\|\phi\|_{\mathcal{L}^2(\Omega)} \\ & - \|u_j\|_{\mathcal{L}^2(\Omega)}\|\rho_{\hat{\gamma}_j}\|_{\mathcal{L}^2(\Omega)} + \frac{1}{2}\mathrm{Tr}(-\Delta\hat{\gamma}) + \epsilon_{k_j} \\ \geq & C_{13}\|\phi\|^2_{\mathcal{L}^2(\Omega)} - C_{14}\|\phi\|_{\mathcal{L}^2(\Omega)} + C_{15}, \end{aligned} \tag{4.194}$$

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

$$\begin{aligned} C_{14} &\equiv \|b(\mathbf{r},\{\mathbf{R_1},\cdots,\mathbf{R_M}\})\|_{\mathcal{L}^2(\Omega)} + \sup_{j\in\mathbb{N}}\|\rho_{\hat{\gamma}_j}\|_{\mathcal{L}^2(\Omega)}, \\ C_{15} &\equiv \sup_{j\in\mathbb{N}}\Big\{ - \|u_j\|_{\mathcal{L}^2(\Omega)}\|\rho_{\hat{\gamma}_j}\|_{\mathcal{L}^2(\Omega)} + \frac{1}{2}\mathrm{Tr}(-\Delta\hat{\gamma}_j) + \epsilon_{k_j}\Big\}. \end{aligned}$$

With (4.194), the equi-coercivity of $-S^{j,k_j}(u_j,\phi)$ with respect to the weak topology in $\mathcal{V}$ is proved. ∎

Lemma 4.5.11 *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 [49], Lemma 4.5.9, and Lemma 4.5.10. ∎

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

Lemma 4.5.12 *The family of functionals $\{T^{j,k_j}(u)\}_{j\in\mathbb{N}}$ converges in the Γ-sense, i.e., for $j\to\infty$,*

$$T^{j,k_j}(u) \xrightarrow{\Gamma} T(u) \tag{4.195}$$

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

Proof We begin by showing the lim-inf condition for

$$T^{j,k_j}(u) = B^*_{\mathrm{xc}}(u) + \sup_{\phi\in\mathcal{V}} S^{j,k_j}(u,\phi). \tag{4.196}$$

From Lemma 4.5.11, we have for every $u_j \rightharpoonup u$ in $\mathcal{U}$ and $u \in \mathcal{U}$,

$$\lim_{j\to\infty} \sup_{\phi\in\mathcal{V}} S^{j,k_j}(u_j,\phi) = \sup_{\phi\in\mathcal{V}} S(u,\phi). \tag{4.197}$$

In addition, $B^*_{\text{xc}}(u)$ is weakly lower semi-continuous, see [17]. 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 (B.6) in the appendix B, the continuity of the functional $B^*_{\text{xc}}(u)$ in $\mathcal{U}$ can be established through Fatou's Lemma:

$$\lim_{j\to\infty} B^*_{\text{xc}}(u_j) = B^*_{\text{xc}}(u). \tag{4.198}$$

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 4.5.13 *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 [17],

$$B^*_{\text{xc}}(u) = \int_\Omega h^*\big(u(\mathbf{r})\big)\,\mathrm{d}\mathbf{r}, \tag{4.199}$$

where $h^*(x) : \mathbb{R} \to \mathbb{R}$ is the Legendre transform of $(-h(t))$ from equation (4.12). Using the bounds from equation (B.6) in Appendix B, there exist real constants $C_{16} > 0$ and C_{17} such that

$$B^*_{\text{xc}}(u) \geq C_{16}\|u\|^4_{\mathcal{U}} - C_{17}|\Omega|. \tag{4.200}$$

The estimate (4.200) implies natural bounds from below on the functional T^{j,k_j},

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

where $\hat{\phi} = 0$ is a test function in $\mathcal{V}$, λ_{LB} denotes the lower bound of the binning interval $[\lambda_{\mathrm{LB}}, \lambda_{\mathrm{UB}}]$ 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

$$\lambda_{\mathrm{LB}} = \lambda_1^{H^j(\hat{\phi},u)} + C_j. \tag{4.201}$$

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,

$$\begin{aligned} \lambda_1^{H^j(\hat{\phi},u)} &= \Big\langle H^j(\hat{\phi},u)\xi_1^{H^j(\hat{\phi},u)}, \xi_1^{H^j(\hat{\phi},u)} \Big\rangle \\ &\geq \|\nabla\xi_1^{H^j(\hat{\phi},u)}\|^2_{\mathcal{L}^2(\Omega)} - \|u\|_{\mathcal{L}^2(\Omega)} \\ &\geq -\|u\|_{\mathcal{L}^2(\Omega)}. \end{aligned} \tag{4.202}$$

Using the inequality (4.202), we can bound $T^{j,k_j}(u)$ from below by a coercive functional which is independent of j and k_j:

$$\begin{aligned} T^{j,k_j}(u) &\geq B^*_{\mathrm{xc}}(u) - N\|\hat{\phi} - u\|_{\mathcal{L}^2(\Omega)} \\ &\geq C_{16}\|u\|^4_{\mathcal{U}} - N\|u\|^2_{\mathcal{U}}. \end{aligned} \tag{4.203}$$

In the limit $\|u\|_{\mathcal{U}} \to \infty$, the term $C_{16}\|u\|_{\mathcal{U}}^4$ dominates, so we have $T^{j,k_j}(u) \to \infty$. Thus 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 KS energy functionals converges to the full KS ground-state energy:*

$$\lim_{j\to\infty} \inf_{u\in\mathcal{U}} T^{j,k_j}(u) = \inf_{u\in\mathcal{U}} T(u) = \epsilon_0. \tag{4.204}$$

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

$$\lim_{j\to\infty} \inf_{\mathcal{U}_j} \sup_{\mathcal{V}_j} \inf_{\mathcal{K}_{N,k_j}^{Hj(\phi,u)}} L(u,\phi,\gamma) = \inf_{\mathcal{U}} \sup_{\mathcal{V}} \inf_{\mathcal{K}_N^{H(\phi,u)}} L(u,\phi,\gamma) = \epsilon_0^{\mathrm{REKS}}. \tag{4.205}$$

Proof This is proven using Theorem 7.8 in [49], Lemma 4.5.12 and Lemma 4.5.13. ∎

Chapter 5

Binning in one dimension, a model problem

We now test the efficiency of the spectral binning scheme on a one-dimensional benchmark problem proposed by Cervera *et al.* [18]. Specifically, we 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:

$$V(r) = -\sum_{i\in\mathbb{Z}} \frac{\alpha}{\sqrt{2\pi}\beta} \exp\left(-\frac{(r-R_i)^2}{2\beta^2}\right). \tag{5.1}$$

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

$$H\psi_i = \Big(-\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}r^2} + V(r)\Big)\psi_i = \epsilon_i\psi_i. \tag{5.2}$$

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, $\alpha = 10$, $\beta = 0.45$, and an insulating chain, $\alpha = 100$, $\beta = 0.3$.

The flowchart of the binning algorithm as used in calculations is as follows:

do *Find an initial guess to* $[\lambda_{\mathrm{LB}}, \lambda_{\mathrm{UB}}]$*;*

Perform a LDL^T *decomposition of* $H^j - \lambda_{\mathrm{LB}}\mathcal{I}^j$ *and* $H^j - \lambda_{\mathrm{UB}}\mathcal{I}^j$*;*

Find $\mathcal{N}_-(H^j - \lambda_{\mathrm{LB}}\mathcal{I}^j)$ *and* $\mathcal{N}_-(H^j - \lambda_{\mathrm{UB}}\mathcal{I}^j)$*;*

if $\mathcal{N}_-(H^j - \lambda_{\mathrm{LB}}\mathcal{I}^j) > 0$*;*

then

Decrease λ_{LB} until $\mathcal{N}_-(H^j - \lambda_{\mathrm{LB}}\mathcal{I}^j) = 0$.

end

if $\mathcal{N}_-(H^j - \lambda_{\mathrm{UB}}\mathcal{I}^j) < N$*;*

then

Increase λ_{UB} until $\mathcal{N}_-(H^j - \lambda_{\mathrm{UB}}) > N$;

else

Use bisection to decrease λ_{UB} so that $\mathcal{N}_-(H^j - \lambda_{\mathrm{UB}}\mathcal{I}^j) = N + \epsilon_N$ with $\epsilon_N \in \mathbb{N}_{>0}$;

end

do *Partition* $[\lambda_{\mathrm{LB}}, \lambda_{\mathrm{UB}}]$ *into* k *intervals with end points* $\{t_0^k, t_1^k, \ldots, t_k^k\}$, $\lambda_{\mathrm{LB}} = t_0^k$ *and* $\lambda_{\mathrm{UB}} = t_k^k$*;*

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* $0 \leq c - q^k \leq 1$ *and* $\sum_{q=1}^{k} c_q^k n_q^{k,j} = N$.

Algorithm 1: Spectral binning.

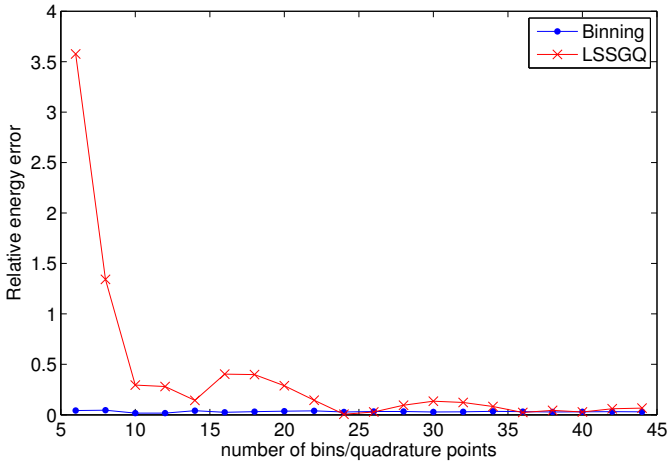

Figure 5.1: Linear chain of M atoms with N electrons [18]. Metal: $\alpha = 10$, $\beta = 0.45$.

A system of $1,000$ atoms and $4,000$ electrons with periodic boundary conditions is discretized using an 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 $[\lambda_{LB}, \lambda_{UB}]$. We use an interior-point method to perform the minimization of (4.110) with respect to the spectral binning coefficients $\{c_q^k\}_{q=1}^k$ subject to the constraints in equation (4.111).

The convergence of the band energies of a metallic and an insulating system calculated using spectral binning and linear-scaling spectral Gauss quadratures (LSSGQ) with a small temperature [73] is shown in Figs. 5.1 and 5.2. We recall that LSSGQ is a linear-scaling method based on polynomial approximations of the Fermi-Dirac distribution (3.6) and the use of associated Gauss quadrature rules. We see that spectral binning outperforms LSSGQ and exhibits comparatively much better accuracy and rate of convergence. The comparison can be made increasingly favorable to binning by further reducing the temperature, since LSSGQ relies on smoothness, whereas binning does not.

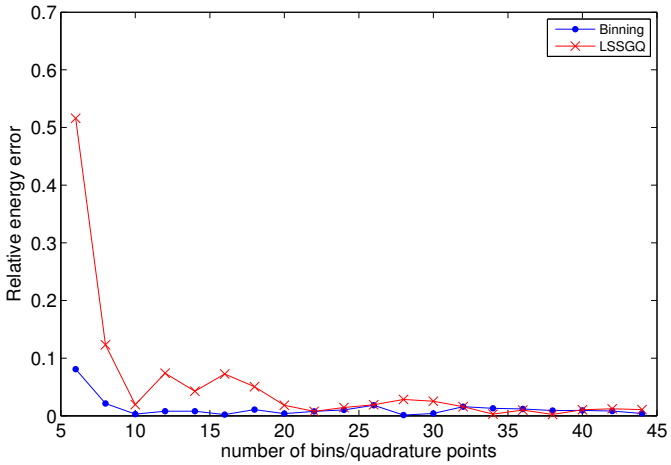

Figure 5.2: Linear chain of M atoms with N electrons [18]. Insulator: $\alpha = 100$, $\beta = 0.3$.

5.1 Discussion

The number of bins required for a given accuracy is independent of the spectrum width $\Delta\lambda$, and therefore, is independent of the spatial discretization, whereas in spectral discretizations using polynomial or rational functions, the number of spectral basis required for a given accuracy grows as the $\Delta\lambda$ increases. This property is advantagous in all-electron calculations or hard pseudopotentials where we need very fine spatial discretizations. The preceding numerical experiments bode well for a general implementation of spectral binning. We note, however, that in attempting such a general implementation, a difficulty that is immediately encountered is that the exchange-correlation functionals that are commonly used in practice are a function of the local electron density. In the context of spectral binning, the electron density $\rho(\mathbf{r})$ is given by

$$\begin{aligned}
\rho_\gamma(\mathbf{r}_0) = \gamma(\mathbf{r}_0, \mathbf{r}_0) = \langle \mathbf{r}_0, \gamma \mathbf{r}_0 \rangle &= \sum_{q=1}^{k} c_q^k \langle \mathbf{r}_0, s_{t_q^k}(H)\mathbf{r}_0 \rangle \\
&= \sum_{q=1}^{k} c_q^k \sum_{p=1}^{\infty} s_{t_q^k}(\lambda_p) \langle \mathbf{r}_0, \xi_p \rangle \langle \xi_p, \mathbf{r}_0 \rangle = \sum_{q=1}^{k} c_q^k \sum_{p=1}^{\infty} s_{t_q^k}(\lambda_p) |\xi_p(\mathbf{r}_0)|^2 \\
&= \sum_{q=1}^{k} c_q^k \sum_{p=1}^{\infty} s_{t_q^k}(\lambda_p) \left| \sum_{m=1}^{\infty} b_m^p e_m(\mathbf{r}_0) \right|^2 , \qquad (5.3)
\end{aligned}$$

where

$$b_m^p \equiv \langle \xi_p, e_m \rangle, \qquad s_{t_q^k}(\lambda_p) \equiv \begin{cases} 1, & \text{if } t_q^k \leq \lambda_p \leq t_{q+1}^k, \\ 0, & \text{otherwise,} \end{cases} \tag{5.4}$$

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 [73], equation (5.3) can be written as

$$\sum_{p=1}^{\infty} s_{t_q^k}(\lambda_p) \left| \sum_{m=1}^{\infty} b_m^p e_m(\mathbf{r}_0) \right|^2 = \int_{\sigma(H)} s_{t_q^k}(\lambda) \, \mathrm{d}\mu_{(\eta_{\mathbf{r}_0}, \eta_{\mathbf{r}_0})} \tag{5.5}$$

and

$$\rho(\mathbf{r}_0) = \sum_{q=1}^{k} c_q^k \int_{\sigma(H)} s_{t_q^k}(\lambda) \, \mathrm{d}\mu_{(\eta_{\mathbf{r}_0}, \eta_{\mathbf{r}_0})}, \tag{5.6}$$

where

$$\eta_{\mathbf{r}_0}(\mathbf{r}) = \sum_{p=1}^{\infty} e_p(\mathbf{r}_0) e_p(\mathbf{r}). \tag{5.7}$$

Thus, the evaluation of the electron density using spectral binning requires the ability to evaluate the quantity $\langle \eta_{\mathbf{r}_0}, s(H)\eta_{\mathbf{r}_0} \rangle$. The efficient evaluation of this quantity without polynomial or rational approximations remains an open problem. This suggests expressing the exchange correlation function in terms of the density matrix directly, which constitutes a natural—but heretofore unexplored—modeling paradigm worthy of further future consideration.

Chapter 6

Conclusion

My PhD work is focused on the approximation methods of the density matrix. The basis of density matrix methods lies in the commutativity between the density matrix and the Kohn-Sham Hamiltonian matrix, i.e., the density matrix can be written as a matrix function of the Kohn-Sham Hamiltonian. The rise of linear-scaling density functional theory methods led to the applications matrix function approximations to density functional theory. A good reference that describes the various ways to approximate a matrix function is the book by Higham [32]. To my knowledge, there has been at least one paper published in the linear-scaling density functional theory literature using the approaches discussed by Higham: from polynomial based approximations using spectral Gauss quadratures to rational function approximations. However, I think there is still room in adapting the implementation of existing approximiations to better suit the architecture of newest supercomputers.

I would like to summarize a couple insights I learned during my PhD regarding the linear scaling spectral Gauss quadrature (LSSGQ) method [72]. First and foremost, LSSGQ requires that the system to be discretized using an orthonormal basis. The requirement derives from the need to compute the trace of the product between the density matrix and the Kohn-Sham density matrix. The trace of a self-adjoint operator is invariant with respect to any orthonormal basis [63]. This requirement rules out the possibility of using conventional finite element methods. One can try to use techniques such as mass-lumping, however it is unclear how the errors introduced by mass-lumping would affect the accuracy of density

functional theory calculations. In addition to the requirement of an orthonormal basis, the Hamiltonian needs to have a sparse representation in the chosen basis. This requirement immediately rules out plane-wave basis. Secondly, the number of spectral quadratures required for a given relative error decreases as the relative Fermi energy, $\hat{\lambda}_{\mathrm{f}}$ increases [70]:

$$\hat{\lambda}_{\mathrm{f}} = \frac{\lambda_{\mathrm{f}}}{\lambda_{\mathrm{max}} - \lambda_{\mathrm{min}}}.$$

To increase the relative Fermi energy, one possibility is to use filtering tools such as Chebyshev filtering.

The other density matrix approximation I investigated during my PhD is spectral binning. Spectral binning is extremely efficient at representing the zero-temperature density matrix function. However, it is unclear how one can extract the electron density at linear-scaling computational cost. It is sufficient to find the electron density if one can compute projections of the matrix sign function at $\mathcal{O}(1)$ cost. Higham [32] suggested several ways in which projections of matrix sign function can be computed:

1. Rational approximations of the square-root function

2. Iterative approximations of the matrix sign-function

The rational approximations of the square-root function bears similarity to the rational approximations such as the pole-expansions [45], one has to investigate whether the rational approximations of the square-root function is more efficient than the pole-expansions. However, it appears that the bottle-neck to rational approximations of the density matrix is not the number of rational expansions, but the computation of each of the rational matrix function of the Kohn-Sham Hamiltonian matrix. The computation of each rational function of the Hamiltonian matrix involves a LDL^T decomposition, which has limitations on parallelizability beyond 10,000 processors. Another potential problem facing spectral binning is the convergence of the zero-temperature density matrix when the Hamiltonian matrix exhibits degeneracy near the Fermi energy. In order to numerically verify this problem, one has to

decide on how to compute the electron density using spectral binning, which would be the focus of future work.

In this thesis, we developed a variational framework to rigorously verify spectral discretizations of the density operator in Kohn-Sham density functional theory. We have proven convergence of both spacial and spectral binning discretizations to the Kohn-Sham ground state energy using our variational framework. Our result is significant because we have been able to show the convergence of spectral binning discretizations to the density operator related to the *non-linear* Kohn-Sham eigenvalue problem, whereas other proofs in literature only show convergence for a *linear* eigenvalue problem. This framework can be extended to prove other types of spectral discretizations such as polynomials and rational functions. We can also include nonzero electronic temperature into our variational framework by simply adding an entropy term to the Kohn-Sham ground state energy. Most importantly, our variational framework can be used to justify the convergence of the self-consistent scheme in Kohn-Sham density functional theory, which has been adopted in calculations without mathematical verification. This variational framework enables us to rigorously verify the linear scaling implementations of Kohn-Sham density functional theory.

Appendix A

Orbital formulation of KSDFT

The KS problem [55] constitutes the minimization of the functional

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

over

$$\left\{ \{\psi_i\} \in \mathcal{V}^N : \langle \psi_i, \psi_j \rangle = \delta_{ij} \right\}, \tag{A.2}$$

where E_{H}, E_{ext}, E_{ZZ}, and E_{xc} are given by (4.9), (4.10), (4.11), and (4.12), 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

$$\left(-\frac{1}{2}\Delta + V \right) \psi = \lambda \psi, \tag{A.3}$$

where

$$V\big(\rho(\mathbf{r}), \mathbf{r}\big) = \int_\Omega \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \, \mathrm{d}\mathbf{r}' + \sum_{1 \le I \le M} \frac{Z_I}{|\mathbf{R}_I - \mathbf{r}|} + h'(\rho(\mathbf{r})). \tag{A.4}$$

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

$$\gamma = \sum_{1 \leq i \leq N} \psi_i \otimes \psi_i \tag{A.5}$$

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., [17]), we have that every functional in $\Gamma(\mathcal{V})$ is convex and lower semi-continuous. Further, we have the following statementi (see Proposition 3.1 in [17]).

Proposition B.0.1 *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

$$F^* = \sup_{u \in \mathcal{T}} \big\{ \langle u^*, u \rangle - F(u) \big\}. \tag{B.1}$$

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 [17]).

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

$$\begin{aligned} -E_{\mathrm{xc}}(\rho_\gamma) &= \sup_{u \in \mathcal{L}^{r'}(\Omega)} \{\langle u, \rho_\gamma \rangle - B_{\mathrm{xc}}(u)^*\} \\ &= - \inf_{u \in \mathcal{L}^{r'}(\Omega)} \{B^*_{\mathrm{xc}} - \langle u, \rho_\gamma \rangle\}, \end{aligned} \tag{B.2}$$

where

$$B^*_{\mathrm{xc}}(u) = (-E_{\mathrm{xc}}(\rho))^* \tag{B.3}$$

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 (4.25).

From Proposition 2.1 in [17], we know that

$$B^*_{\mathrm{xc}}(u) = \int_\Omega h^*(u)\, \mathrm{d}\mathbf{r}, \tag{B.4}$$

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

$$C_1 |t|^{\frac{4}{3}} + C_2 \leq -h(t) \leq C_3 |t|^{\frac{4}{3}} + C_4 \tag{B.5}$$

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

$$C_{18} |x|^4 + C_{19} \leq h^*(x) \leq C_{16} |x|^4 + C_{17} \tag{B.6}$$

for $h^*(x)$.

Bibliography

[1] Robert A. Adams and John J. F. Fournier. *Sobolev Spaces*. Academic Press, 2003.

[2] Arnaud Anantharaman and Eric Cancès. Existence of minimizers for KohnSham models in quantum chemistry. *Annales de l'Institut Henri Poincare (C) Non Linear Analysis*, 26(6):2425–2455, November 2009.

[3] E. Antončík. On the approximate formulation of the orthogonalized plane-wave method. *Czechoslovak Journal of Physics*, 10(1):22–27, January 1960.

[4] Rickard Armiento, Boris Kozinsky, Marco Fornari, and Gerbrand Ceder. Screening for high-performance piezoelectrics using high-throughput density functional theory. *Physical Review B*, 84(1):014103, July 2011.

[5] William Arveson. *Ten Lectures on Operator Algebras, Issue 55*. American Mathematical Soc., 1984.

[6] Roi Baer and Martin Head-Gordon. Chebyshev expansion methods for electronic structure calculations on large molecular systems. *The Journal of Chemical Physics*, 107(23):10003, December 1997.

[7] A. D. Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical Review A*, 38(6):3098–3100, September 1988.

[8] Axel D. Becke. A new mixing of HartreeFock and local density-functional theories. *The Journal of Chemical Physics*, 98(2):1372, January 1993.

[9] F.A. Berezin and Mikhail Aleksandrovich Shubin. *The Schrödinger Equation*. Springer Science & Business Media, 1991.

[10] D R Bowler, T Miyazaki, and M J Gillan. Recent progress in linear scaling ab initio electronic structure techniques. *Journal of Physics: Condensed Matter*, 14(11):2781–2798, March 2002.

[11] Haim Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, volume 2010. Springer, 2010.

[12] Eric Cancès, Mireille Defranceschi, Werner Kutzelnigg, Claude Le Bris, and Yvon Maday. *Special Volume, Computational Chemistry*, volume 10 of *Handbook of Numerical Analysis*. Elsevier, 2003.

[13] D. M. Ceperley. Ground State of the Electron Gas by a Stochastic Method. *Physical Review Letters*, 45(7):566–569, August 1980.

[14] M. L. Cohen and V. Heine. The fitting of pseudopotentials to experimental data and their subsequent application. *Solid State Physics*, 24:37, 1970.

[15] P. A. M. Dirac. Note on Exchange Phenomena in the Thomas Atom. *Mathematical Proceedings of the Cambridge Philosophical Society*, 26(03):376, October 1930.

[16] Jeff W. Doak and C. Wolverton. Coherent and incoherent phase stabilities of thermoelectric rocksalt IV-VI semiconductor alloys. *Physical Review B*, 86(14):144202, October 2012.

[17] Ivar Ekeland and Roger Témam. *Convex Analysis and Variational Problems, Volume 1*. Society for Industrial and Applied Mathematics, 1999.

[18] C. García-Cervera, Jianfeng Lu, Yulin Xuan, and Weinan E. Linear-scaling subspace-iteration algorithm with optimally localized nonorthogonal wave functions for Kohn-Sham density functional theory. *Physical Review B*, 79(11):115110, March 2009.

[19] Vikram Gavini, Jaroslaw Knap, Kaushik Bhattacharya, and Michael Ortiz. Non-periodic finite-element formulation of orbital-free density functional theory. *Journal of the Mechanics and Physics of Solids*, 55(4):669–696, April 2007.

[20] S. Goedecker and L. Colombo. Efficient Linear Scaling Algorithm for Tight-Binding Molecular Dynamics. *Physical Review Letters*, 73(1):122–125, July 1994.

[21] S. Goedecker and O.V. Ivanov. Linear scaling solution of the Coulomb problem using wavelets. *Solid State Communications*, 105(11):665–669, March 1998.

[22] S. Goedecker and M. Teter. Tight-binding electronic-structure calculations and tight-binding molecular dynamics with localized orbitals. *Physical Review B*, 51(15):9455–9464, April 1995.

[23] Stefan Goedecker. Linear scaling electronic structure methods. *Reviews of Modern Physics*, 71(4):1085–1123, July 1999.

[24] Gene H. Golub and Gérard Meurant. *Matrices, Moments and Quadrature with Applications*. Princeton University Press, 2009.

[25] D. Hamann, M. Schlüter, and C. Chiang. Norm-Conserving Pseudopotentials. *Physical Review Letters*, 43(20):1494–1497, November 1979.

[26] Walter A. Harrison. *Electronic Structure and the Properties of Solids: The Physics of the Chemical Bond*. Courier Corporation, 2012.

[27] Geoffroy Hautier, Shyue Ping Ong, Anubhav Jain, Charles J. Moore, and Gerbrand Ceder. Accuracy of density functional theory in predicting formation energies of ternary oxides from binary oxides and its implication on phase stability. *Physical Review B*, 85(15):155208, April 2012.

[28] H. Hellmann. A New Approximation Method in the Problem of Many Electrons. *The Journal of Chemical Physics*, 3(1):61, November 1935.

[29] E. Hernández, M. Gillan, and C. Goringe. Linear-scaling density-functional-theory technique: The density-matrix approach. *Physical Review B*, 53(11):7147–7157, March 1996.

[30] Conyers Herring. A New Method for Calculating Wave Functions in Crystals. *Physical Review*, 57(12):1169–1177, June 1940.

[31] T Hickel, B Grabowski, F Körmann, and J Neugebauer. Advancing density functional theory to finite temperatures: methods and applications in steel design. *Journal of physics. Condensed matter : an Institute of Physics journal*, 24(5):053202, February 2012.

[32] Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2008.

[33] P. Hohenberg. Inhomogeneous Electron Gas. *Physical Review*, 136(3B):B864–B871, November 1964.

[34] Sohrab Ismail-Beigi and T. Arias. Locality of the Density Matrix in Metals, Semiconductors, and Insulators. *Physical Review Letters*, 82(10):2127–2130, March 1999.

[35] Sohrab Ismail-Beigi and T.A. Arias. New algebraic formulation of density functional calculation. *Computer Physics Communications*, 128(1-2):1–45, June 2000.

[36] Jürgen Jost and Xianqing Li-Jost. *Calculus of Variations*. Cambridge University Press, 1998.

[37] Leonard Kleinman and D. M. Bylander. Efficacious form for model pseudopotentials. *Physical Review Letters*, 48:1425–1428, 1982.

[38] W. Kohn and L. J. Sham. Self-Consistent Equations Including Exchange and Correlation Effects. *Physical Review*, 140(4A):A1133–A1138, November 1965.

[39] Cornelius Lanczos. An iterative method for the solution of the eigenvalue problem of linear differential and integral. *Journal of Research of the National Bureau of Standards*, 45(4), 1950.

[40] David Langreth and M. Mehl. Easily Implementable Nonlocal Exchange-Correlation Energy Functional. *Physical Review Letters*, 47(6):446–450, August 1981.

[41] Mel Levy. Electron densities in search of Hamiltonians. *Physical Review A*, 26(3):1200–1208, September 1982.

[42] X.-P. Li, R. Nunes, and David Vanderbilt. Density-matrix electronic-structure method with linear system-size scaling. *Physical Review B*, 47(16):10891–10894, April 1993.

[43] Elliott H. Lieb. Density functionals for coulomb systems. *International Journal of Quantum Chemistry*, 24(3):243–277, September 1983.

[44] Lin Lin, Mohan Chen, Chao Yang, and Lixin He. Accelerating atomic orbital-based electronic structure calculation via pole expansion and selected inversion. *Journal of physics. Condensed matter : an Institute of Physics journal*, 25:295501, 2013.

[45] Lin Lin, Jianfeng Lu, Lexing Ying, and E. Weinan. Pole-Based approximation of the Fermi-Dirac function. *Chinese Annals of Mathematics, Series B*, 30(6):729–742, August 2009.

[46] Norman Henry March. *The Many-body Problem in Quantum Mechanics*. Dover Publications, 1967.

[47] Richard M. Martin. *Electronic Structure: Basic Theory and Practical Methods*. Cambridge University Press, 2004.

[48] Nicola Marzari and David Vanderbilt. Maximally localized generalized Wannier functions for composite energy bands. *Physical Review B*, 56(20):12847–12865, November 1997.

[49] Gianni Dal Maso. *An Introduction to G-Convergence*. Springer, 1993.

[50] Francesco Mauri and Giulia Galli. Electronic-structure calculations and molecular-dynamics simulations with linear system-size scaling. *Physical Review B*, 50(7):4316–4326, August 1994.

[51] Phani Motamarri and Vikram Gavini. A subquadratic-scaling subspace projection method for large-scale Kohn-Sham density functional theory calculations using spectral finite-element discretization. *Physical Review B*, 90:25, June 2014.

[52] Marco Nava, Michele Ceriotti, Chaim Dryzun, and Michele Parrinello. Evaluating functions of positive-definite matrices using colored-noise thermostats. *Physical Review E*, 89(2):023302, February 2014.

[53] Pablo Ordejón, David Drabold, Matthew Grumbach, and Richard Martin. Unconstrained minimization approach for electronic computations that scales linearly with system size. *Physical Review B*, 48(19):14646–14649, November 1993.

[54] Beresford N. Parlett. *The Symmetric Eigenvalue Problem*. SIAM, 1998.

[55] Robert G. Parr and Robert G. Parr Weitao Yang. *Density-Functional Theory of Atoms and Molecules*. Oxford University Press, 1989.

[56] J. P. Perdew and Alex Zunger. Self-interaction correction to density-functional approximations for many-electron systems. *Physical Review B*, 23:5048–5079, 1981.

[57] John Perdew. Accurate Density Functional for the Energy: Real-Space Cutoff of the Gradient Expansion for the Exchange Hole. *Physical Review Letters*, 55(16):1665–1668, October 1985.

[58] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized Gradient Approximation Made Simple. *Physical Review Letters*, 77(18):3865–3868, October 1996.

[59] John P. Perdew and Yue Wang. Accurate and simple analytic representation of the electron-gas correlation energy. *Physical Review B*, 45:13244–13249, 1992.

[60] John P. Perdew and Wang Yue. Accurate and simple density functional for the electronic exchange energy: Generalized gradient approximation. *Physical Review B*, 33(12):8800–8802, June 1986.

[61] James Phillips and Leonard Kleinman. New Method for Calculating Wave Functions in Crystals and Molecules. *Physical Review*, 116(2):287–294, October 1959.

[62] Warren E. Pickett. Pseudopotential methods in condensed matter applications. *Computer Physics Reports*, 9(3):115–197, April 1989.

[63] Michael Reed and Barry Simon. Reed Simon - Vol. 1 Methods of mathematical physics - Functional analysis.pdf, 1981.

[64] Anindya Roy, Joseph W. Bennett, Karin M. Rabe, and David Vanderbilt. Half-Heusler Semiconductors as Piezoelectrics. *Physical Review Letters*, 109(3):037602, July 2012.

[65] Walter Rudin. *Functional Analysis*. McGraw-Hill, 1991.

[66] James E. Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and C. Wolverton. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM*, 65(11):1501–1509, September 2013.

[67] S. Sandlöbes, M. Friák, S. Zaefferer, a. Dick, S. Yi, D. Letzig, Z. Pei, L.-F. Zhu, J. Neugebauer, and D. Raabe. The relation between ductility and stacking fault energies in Mg and MgY alloys. *Acta Materialia*, 60(6-7):3011–3021, April 2012.

[68] Grady Schofield, James R. Chelikowsky, and Yousef Saad. A spectrum slicing method for the KohnSham problem. *Computer Physics Communications*, 183(3):497–505, March 2012.

[69] R. Shankar. *Principles of Quantum Mechanics*. Springer, 1994.

[70] Phanish Suryanarayana. On spectral quadrature for linear-scaling Density Functional Theory. *Chemical Physics Letters*, 584:182–187, October 2013.

[71] Phanish Suryanarayana. Optimized purification for density matrix calculation. *Chemical Physics Letters*, 555:291–295, January 2013.

[72] Phanish Suryanarayana, K Bhattacharya, and M Ortiz. Coarse-graining KohnSham Density Functional Theory. *Journal of the Mechanics and ...*, 2011, 2013.

[73] Phanish Suryanarayana, Kaushik Bhattacharya, and Michael Ortiz. Coarse-graining KohnSham Density Functional Theory. *Journal of the Mechanics and Physics of Solids*, 61(1):38–60, January 2013.

[74] Phanish Suryanarayana, Vikram Gavini, Thomas Blesgen, Kaushik Bhattacharya, and Michael Ortiz. Non-periodic finite-element formulation of KohnSham density functional theory. *Journal of the Mechanics and Physics of Solids*, 58(2):256–280, February 2010.

[75] 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 Series*, 4(23):138–142, April 1852.

[76] B. Tong and L. Sham. Application of a Self-Consistent Scheme Including Exchange and Correlation Effects to Atoms. *Physical Review*, 144(1):1–4, April 1966.

[77] Dallas R. Trinkle, Joseph A. Yasi, and Louis G. Hector. *Predicting Mg Strength from First-principles: Solid-solution Strengthening, Softening, and Cross-slip*. John Wiley and Sons, Inc., 2011.

[78] N. Troullier and José Luriaas Martins. Efficient pseudopotentials for plane-wave calculations. *Physical Review B*, 43(3):1993–2006, January 1991.

[79] David Vanderbilt. Soft self-consistent pseudopotentials in a generalized eigenvalue formalism. *Physical Review B*, 41(11):7892–7895, April 1990.

[80] E. J. M. Veling. Lower bounds for the infimum of the spectrum of the Schrödinger operator in RN and the Sobolev inequalities. *Journal of inequalities in pure and applied mathematics*, 3(4), 2002.

[81] S. H. Vosko, L. Wilk, and M. Nusair. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Canadian Journal of Physics*, 58(8):1200–1211, August 1980.

[82] Richard Wheeden and Antoni Zygmund. *Measure and Integral: An Introduction to Real Analysis*, volume 1977. CRC Press, 1977.