Chapter 1

Introduction

1.1 Overview

The arrangement of the electrons characterizes the microscopic structure of molecules and systems in condensed phases in chemistry, biology, and material science. The arrangement of the electrons is described by the electronic structure theory. The various forms of the electronic structure theory differ by orders of magnitude from each other in terms of accuracy and efficiency. Among the different formalisms of the electronic structure theory, Kohn-Sham density functional theory (KSDFT) achieves the best compromise between accuracy and efficiency, and is by far the most widely used electronic structure theory. Nonetheless, the computational cost of KSDFT still increases rapidly with respect to the number of electrons in the system, which hinders the application of KSDFT to systems of large size. Reducing the computational cost of KSDFT requires combined knowledge of mathematics, physics and computer science. Part I of this dissertation explores the mathematical properties of KSDFT, and develops an accurate and efficient algorithm for applying KSDFT to systems of large scale.

The scale of the system, *i.e.* the number of the electrons is a crucial parameter in many applications in chemistry, biology, and material science. It is desirable to have the same number of electrons in the numerical simulation as that in the real system. However, even a water droplet contains more than 10^{20} electrons. This overwhelmingly large magnitude is out of the scope of any of the existing simulation technique, and samples of smaller size has to be used in practice. The small sample size introduces a systematic error to the numerical simulation, called the size effect, which is non-negligible in many applications. For example, the diffusion constant of a polymer chain is underestimated by $2 \sim 3$ times due to the size effect [79]; Crack propagation [1] involves tens of thousands to millions of atoms and electrons by its nature, and cannot be observed with samples of smaller sizes. Systems with millions of electrons are usually treated by classical mechanics with empirical potential. The empirical potential energy surfaces have achieved success in describing a large class of phenomena provided that the empirical parameters are carefully optimized. On the other hand, the potential energy surfaces directly generated from quantum mechanics, such as from KSDFT, have the advantage that the computational result depends only on a small number of universal parameters, including atomic species, atomic positions, and a few parameters in the energy functional which do not depend on the specific setup of the system. As a result, simulations with potential energy surfaces generated from KSDFT are called "first principle" simulations, and are capable of treating bond-forming, bond-breaking, cracking, and other complicated chemical and mechanical processes without extra tuning of the parameters. It is thus desirable to directly apply KSDFT to study systems consisting millions of electrons, but this is far beyond the current capability of the KSDFT. In the standard methods for solving KSDFT, the computational cost increases as $\mathcal{O}(N^3)$ where N is the number of electrons in the system. Although the standard algorithm has been highly optimized in the past two decades to reduce the computational cost in practice [94,145,182,256], the cubic scaling still limits the application of KSDFT to systems with at most tens of thousands of electrons.

Various efforts have been devoted to reduce the cubic scaling of KSDFT in the past two decades. The major breakthrough is achieved by the algorithms with linear scaling, *i.e.* $\mathcal{O}(N)$ scaling [88,89,91,98,99,101,102,154,184,193,229,252]. Such linear scaling algorithms relies on the "nearsightedness" property of the electrons [142,212], which means that the density perturbation induced by a local change in the external potential dies off exponentially with respect to the distance from the place where the perturbation was applied. According to DFT the ground state energy is a functional of the density, then the effect of a local perturbation on the density is also local because of nearsightedness, and the energy should not have a very long-range dependence on the density. The nearsightedness property allows one to divide the entire system into many pieces. The size of each piece has fixed size, and the total number of pieces is proportional to the number of electrons in the entire system. The computational cost for solving KSDFT in each piece is fixed, and the total computational cost is therefore proportional to the number of electrons in the entire system. Therefore, $\mathcal{O}(N)$ scaling is achieved.

The nearsightedness property is not valid for all systems, but is only valid for a class of materials called insulating systems, including sulfur, glass, paper, large organic molecules such as DNA and protein, most of the common salts and oxides, to name a few. The nearsightedness property is violated in metallic systems, namely the density perturbation induced by a local change in the external potential exhibits algebraic and oscillatory decay (called the Friedel oscillation [87]) with respect to the distance from the place where the perturbation was applied. Two thirds of the elements in the periodic table can directly form metallic systems, such as aluminum, lithium, copper, and iron. Non-metallic elements can also form metallic systems, *e.g.* graphene and carbon nanotube which only consists of carbon. Due to the violation of the nearsightedness property, the $\mathcal{O}(N)$ scaling algorithm is not directly applicable to metallic systems. The nearsightedness property can be recovered by introducing an artificial finite temperature in the system, and the decay rate of the density perturbation induced by a local change in the external potential becomes exponential again with the exponent depending on the artificial temperature [100, 140].

However, it is not easy to take into account the nearsightedness in practical algorithms. First, the nearsightedness is not so precisely defined in practice, and particularly the nearsightedness decay length is difficult to be predicted in advance. Second, although the accuracy of the linear scaling methods can be systematically improved by increasing the size of each piece (usually characterized by a truncation radius R), the truncation radius R can still be quite large if high numerical accuracy is to be achieved. This is especially the case for metallic system where the truncation radius depends explicitly on the artificial temperature. The magnitude of the artificial temperature should be chosen carefully in order to balance the accuracy and the computational cost. Third, it is not a trivial step to implement the nearsightedness if the nearsightedness is imposed as a constraint on the Kohn-Sham orbitals or the Kohn-Sham single particle density matrix. As a result, the $\mathcal{O}(N^3)$ scaling methods are still the best uniform solution for applying KSDFT to insulating systems and to metallic systems. The important question remains open: whether it is possible to improve the $\mathcal{O}(N^3)$ scaling uniformly for all the systems without encountering the difficulties in nearsightedness algorithms?

This open question is positively answered in Part I of this dissertation. Instead of using the nearsightedness property, this dissertation focuses on the mathematical properties of KSDFT that are uniformly valid for insulating systems and metallic systems, at low and at high temperature. As a result, this dissertation develops algorithms with improved computational scaling over $\mathcal{O}(N^3)$ scaling methods for all systems. To be more specific, the amount of improvement depends on the dimension of the system. The computational cost of the present algorithm is $\mathcal{O}(N)$ for one dimensional systems, $\mathcal{O}(N^{1.5})$ for two-dimensional systems, and $\mathcal{O}(N^2)$ for three-dimensional systems. Furthermore, the present algorithm can be combined with the nearsightedness property, and achieves $\mathcal{O}(N)$ scaling at all dimensions for insulating systems and for metallic system at high temperature.

This chapter provides the minimum amount of prerequisite knowledge for Part I of this dissertation. The rest of this chapter is organized as follows: Section 1.2 briefly introduces the quantum many body problem, and the electronic structure problem with the Bohr-Oppenheimer approximation, followed by Section 1.3 for the basic components of the Kohn-Sham density functional theory. The pseudopotential framework for KSDFT is introduced in Section 1.4. In Section 1.5 the mathematical properties of the KSDFT which are essential for the new method developed in Part I of this dissertation are discussed. Section 1.6 reviews the existing methods and the most widely used software packages for solving KSDFT. Finally Section 1.7 outlines the various components of the new method that will be discussed in detail in the rest of the Chapters in Part I.

1.2 Quantum many body problem and electronic structure

The microscopic properties of electrons in chemistry, biology and material science are accurately described by the many body Hamiltonian of the Schrödinger equation. The many body Hamiltonian associated with a system with N_{nuc} atoms and N electrons is

$$H = \sum_{I=1}^{N_{\text{nuc}}} \frac{P_I^2}{2M_I} + \sum_{i=1}^{N} \frac{p_i^2}{2} + V(R_1, \ldots, R_{N_{\text{nuc}}}, x_1, \ldots, x_N). \tag{1.1}$$

Atomic units are used throughout this dissertation. Namely, without further specification, the unit of energy is Hartree, the unit of length is Bohr, the unit of mass is the electron mass m_e, the unit of charge is the electron charge e, and the Planck constant $\hbar$ equals to 1. Moreover, M_I is the mass of the I-th nucleus, R_I is the position of the I-th nucleus, and x_i is the position of the i-th electron. The momentum operator of the nucleus and the electron are denoted by P_I, p_i as $p_i = -i\nabla_{x_i}$, $P_I = -i\nabla_{R_I}$. Spin is neglected at the moment. V is the interaction energy between the nuclei and the electrons, given by

$$V(R_1, \ldots, R_{N_{\mathrm{nuc}}}, x_1, \ldots, x_N) = \frac{1}{2}\sum_{I \neq J} \frac{Z_I Z_J}{|R_I - R_J|} + \frac{1}{2}\sum_{i \neq j} \frac{1}{|x_i - x_j|} - \sum_{i,I} \frac{Z_I}{|x_i - R_I|}. \tag{1.2}$$

The charge of the I-th nucleus is Z_I. The three terms in Eq. (1.2) represent the Coulomb interactions among nuclei-nuclei, electron-electron and nuclei-electron, respectively.

The many body Hamiltonian (1.1) contains all the information of systems, but all the information does not have equal importance in practice. In many cases, the important information is only contained in the most stable state of the system. This most stable state is characterized by the *ground state* of the many body Hamiltonian, *i.e.* the smallest eigenvalue and eigenvector of the many body Hamiltonian.

$$H\Psi(R_1, \ldots, R_{N_{\mathrm{nuc}}}; x_1, \ldots, x_N) = E\Psi(R_1, \ldots, R_{N_{\mathrm{nuc}}}; x_1, \ldots, x_N). \tag{1.3}$$

Eq. (1.3) is referred to as the *quantum many body problem.* E is called the ground state energy of the many body system, and Ψ is called the ground state wavefunction. Ψ should satisfy certain symmetry condition determined by the statistics of both electrons and nuclei. Especially, Ψ is an antisymmetric function with respect to the coordinates for the electrons $(x_1, \cdots, x_N)$. Ψ changes sign if any pair of coordinates

x_i and x_j are interchanged:

$$\begin{aligned}\Psi(&R_1,\ldots,R_{N_{\mathrm{nuc}}};x_1,\ldots,x_i,\ldots,x_j,\ldots,x_N)\\&=-\Psi(R_1,\ldots,R_{N_{\mathrm{nuc}}};x_1,\ldots,x_j,\ldots,x_i,\ldots,x_N),\quad 1\le i<j\le N.\end{aligned}\tag{1.4}$$

The many body problem (1.3) can be analytically solved for a hydrogen atom which contains only one nucleus and one electron. The solution of (1.3) is already much more complicated in an H_2 molecule which contains two electrons and two nuclei. The solution becomes generally intractable for more than 20 particles even with the help of numerical methods and increasingly more powerful computers. The reason for this enormous complexity is that the dimension of the ground state wavefunction is $3(N_{\mathrm{nuc}}+N)$. Even if each spatial coordinate is discretized by 10 points, $10^{3(N_{\mathrm{nuc}}+N)}$ real numbers will be immediately required just to record one state of the system. Although the quantum many body problem is in principle an exact theory, it exhibits exponential complexity and is intractable without further theoretical approximation. The enormous complexity of the quantum many body problem was well summarized by Dirac in 1929 [73]: "The fundamental laws necessary to the mathematical treatment of large parts of physics and the whole of chemistry are thus fully known, and the difficult lies only in the fact that application of these laws leads to equations that are too complex to be solved."

The first step to reduce the complexity of the quantum many body problem is the Born-Oppenheimer approximation [38], which separates the complexity due to the electrons and that due to the nuclei. The mass of the electron is more than a thousand times smaller than the mass of the nuclei of the lightest element in the periodic table, *i.e.* hydrogen. The Born-Oppenheimer approximation recognizes that the electrons should therefore move much faster than the nuclei, and the state of the electrons is "slaved" to the motion of nuclei. More specifically, for fixed nuclei positions $(R_1, \dots, R_{N_{\mathrm{nuc}}})$, the state of the electrons is described by the ground state of the many body Hamiltonian of the electrons:

$$H_e = -\frac{1}{2}\sum_i \Delta_{x_i} + \sum_i V_{\mathrm{ext}}(x_i) + V_{ee}(x_1, \dots, x_N). \tag{1.5}$$

The nuclei-electron interaction V_{ext} and the electron-electron interaction V_{ee} are defined as

$$V_{\mathrm{ext}}(x) = -\sum_I \frac{Z_I}{|x - R_I|}, \quad V_{ee}(x_1, \dots, x_N) = \frac{1}{2}\sum_{i \neq j} \frac{1}{|x_i - x_j|}. \tag{1.6}$$

Compared to Eq. (1.2), the nuclei-nuclei interaction is excluded from H_e, since the nuclei-nuclei interaction

$$V_{nn}(R_1, \dots, R_{N_{\mathrm{nuc}}}) = \frac{1}{2}\sum_{I \neq J} \frac{Z_I Z_J}{|R_I - R_J|}, \tag{1.7}$$

is a constant term for fixed nuclei positions $(R_1, \dots, R_{N_{\mathrm{nuc}}})$.

The ground state of the many body Hamiltonian of the electrons is defined by

$$H_e \Psi_e(r_1, \dots, r_N; R_1, \dots, R_{N_{\mathrm{nuc}}}) = E_e(R_1, \dots, R_{N_{\mathrm{nuc}}}) \Psi_e(r_1, \dots, r_N; R_1, \dots, R_{N_{\mathrm{nuc}}}), \tag{1.8}$$

E_e is called the ground state energy of the electrons. Ψ_e is called the ground state wavefunction of the electrons, and is an antisymmetric function.

The ground state energy $E_e(R_1, \dots, R_{N_{\mathrm{nuc}}})$ has important physical meaning. The ground state energy together with the nuclei-nuclei interaction V_{nn} form the effective inter-atomic potential $V_{\mathrm{eff}}(R_1, \dots, R_{N_{\mathrm{nuc}}}) = E_e(R_1, \dots, R_{N_{\mathrm{nuc}}}) + V_{nn}(R_1, \dots, R_{N_{\mathrm{nuc}}})$. This inter-atomic potential is completely determined by the atomic species and the atomic positions, and has great advantage over the classical inter-atomic potential characterized by empirical parameters. Therefore, the ground state energy E_e

carries most of the information of the arrangements of the electrons. Calculating $E_e(R_1, \ldots, R_{N_{\text{nuc}}})$ from fixed nuclei positions $(R_1, \ldots, R_{N_{\text{nuc}}})$ is called the *electronic structure problem.*

After solving the electronic structure problem, the motions of the nuclei can be approximated by classical mechanics

$$M_I \ddot{R}_i = -\frac{\partial V_{\text{eff}}(R_1, \ldots, R_{N_{\text{nuc}}})}{\partial R}. \tag{1.9}$$

The approximation (1.9) can be improved by more sophisticated techniques such as path integrals formulation [83], which will be discussed in detail in Part II of this dissertation.

From now on we focus on the electronic structure problem, and we drop the subscript e in the ground state energy E_e and in the ground state wavefunction Ψ_e without causing ambiguity.

1.3 Kohn-Sham Density functional theory (KSDFT)

Born-Oppenheimer approximation reduces the quantum many body problem (1.3) to the electronic structure problem. The electronic structure problem still exhibits exponential complexity with respect to the number of electrons N, and it is necessary to make further approximations. Compared to the common acceptance of the Born-Oppenheimer approximation, less agreement is achieved on the approximation of the electronic structure problem. Various electronic structure theories with different accuracy and efficiency have been proposed, including Hartree-Fock [152, 159], configuration interaction [237], coupled cluster [63], Møller-Plesset perturbation theory [187], quantum Monte Carlo [51, 85], and density functional theory [125, 143], to name a few. After decades of development, density functional theory is commonly found to achieve the best compromise between accuracy and efficiency, and has become the most widely used electronic structure theory.

The foundation of the density functional theory is the Hohenberg-Kohn theorem [125]. Hohenberg-Kohn theorem proves that the ground state electron energy E is uniquely determined by the electron density ρ up to a shift of a trivial constant. This dependence is denoted by the density functional $E[\rho]$. Given the N-body wave function Ψ, the electron density is defined as

$$\rho(x) = N \int |\Psi(x, x_2, \ldots, x_N)|^2 \, \mathrm{d}x_2 \cdots \mathrm{d}x_N, \tag{1.10}$$

$\rho(x)$ represents the probability of finding any of the N electrons at point x. The electron density $\rho(x)$ is a function of three coordinates rather than $3N$ coordinates. Therefore density functional theory remarkably reduces the complexity of the electronic structure problem.

If the exact form of the energy functional $E[\rho]$ is known, the ground state energy can be readily obtained by a minimization procedure over a three-dimensional function ρ with respect to the energy functional $E[\rho]$. However, Hohenberg-Kohn theorem only claims the existence of such energy functional without predicting the full detail of its actual form. Furthermore, the energy functional carefully chosen for one system can fail drastically for another system in principle.

The ground-breaking work is provided by Kohn and Sham [143]. Kohn and Sham approximated the energy functional of interacting electrons by an energy functional of non-interacting electrons together with a correction term. The energy functional of non-interacting electrons can be written analytically and contributes to most part of the ground state energy. The remaining correction term, which is called exchange-correlation functional, remains unknown but is relatively easy to be approximated roughly.

The Kohn-Sham density functional theory can be *formally* written down as follows [143]. The rigorous derivation, however, should follow the Levy-Lieb approach of constrained minimization [158]. First, the ground state energy of an interacting inhomogeneous system can be written as [125]

$$E[\rho] = \int V_{\mathrm{ext}}(x)\rho(x)\,\mathrm{d}x + \frac{1}{2}\iint \frac{\rho(x)\rho(y)}{|x-y|}\,\mathrm{d}x\,\mathrm{d}y + G[\rho], \tag{1.11}$$

where the first term (V_{ext}) characterizes the nuclei-electron interaction and the second term gives the electron-electron interaction. $G[\rho]$ is a universal functional of the electron density. The Kohn-Sham density functional theory then approximates $G[\rho]$ as

$$G[\rho] \equiv E_{\mathrm{K}}[\rho] + E_{\mathrm{xc}}[\rho], \tag{1.12}$$

where $E_{\mathrm{K}}[\rho]$ is the kinetic energy of N non-interacting electrons. $E_{\mathrm{xc}}[\rho]$ is defined to be the exchange-correlation energy, which takes into account all the remaining ground state energies that are not represented by the previous terms. The many body wavefunction of N non-interacting electrons takes the form of the Slater determinant

$$\Psi(x_1, \cdots, x_N) = \frac{1}{\sqrt{N!}} \det \begin{pmatrix} \psi_1(x_1) & \cdots & \psi_1(x_N) \\ \vdots & \ddots & \vdots \\ \psi_N(x_1) & \cdots & \psi_N(x_N) \end{pmatrix}, \tag{1.13}$$

where the three-dimensional wavefunctions ψ_i are called the *electron orbitals.* The electron orbitals satisfy the orthonormal condition

$$\int \psi_i(x)^* \psi_j(x)\,\mathrm{d}x = \delta_{ij}. \tag{1.14}$$

The electron density is reconstructed from the electron orbitals according to the relation

$$\rho(x) = \sum_{i=1}^{N} |\psi_i(x)|^2. \tag{1.15}$$

The kinetic energy for the non-interacting electrons is

$$E_{\mathrm{K}}\left[\{\psi_i\}_{i=1}^{N}\right] = \frac{1}{2}\sum_{i=1}^{N}\int |\nabla\psi_i|^2 \,\mathrm{d}x. \tag{1.16}$$

As a result, the Kohn-Sham energy functional is given by

$$E_{\mathrm{KS}}\left[\{\psi_i\}\right] = E_{\mathrm{K}}\left[\{\psi_i\}\right] + \int V_{\mathrm{ext}}(x)\rho(x)\,\mathrm{d}x + \frac{1}{2}\iint \frac{\rho(x)\rho(y)}{|x-y|}\,\mathrm{d}x\,\mathrm{d}y + E_{\mathrm{xc}}[\rho]. \tag{1.17}$$

To find the ground state energy, the energy functional (1.17) should be optimized over all possible electron orbitals $\{\psi_i\}_{i=1}^{N}$, and hence over all possible electron density ρ satisfying $\int \rho(x)\,\mathrm{d}x = N$. On the other hand, if ρ minimizes Kohn-Sham energy (1.17), the corresponding electron orbitals are also determined by the minimization procedure. Therefore the ground state energy of the Kohn-Sham density functional theory depends only on the electron density ρ. This issue will become clearer in Section 1.5.

The choice of the exchange-correlation functional remains unknown. Fortunately, it turned out that even the crudest approximation of exchange-correlation functional, namely the local density approximation (LDA) [52, 206] is often surprisingly accurate for systems with slowly varying charge densities. For example, the bond lengths and bond angles can be predicted by LDA within a few percent for many systems. More sophisticated exchange-correlation functionals such as generalized gradient approximation (GGA) [22, 149, 204], and hybrid exchange-correlation functionals [23, 205] further extend the applicability of KSDFT to a large class of molecules and systems in condensed phase. Without much loss of generality, in the following we will use the LDA form for exchange-correlation functional, *i.e.*

$$E_{\mathrm{xc}}[\rho] = \int \epsilon_{\mathrm{xc}}[\rho(x)]\,\mathrm{d}x. \tag{1.18}$$

To sum up, the Kohn-Sham density functional theory solves the following minimization problem over the electron orbitals $\{\psi_i\}$.

$$\begin{aligned}
E_{\mathrm{KS}} = \min_{\{\psi_i\}} \frac{1}{2}\sum_{i=1}^{N} \int |\nabla\psi_i|^2 \,\mathrm{d}x + \int V_{\mathrm{ext}}(x)\rho(x)\,\mathrm{d}x + \frac{1}{2}\iint \frac{\rho(x)\rho(y)}{|x-y|}\,\mathrm{d}x\,\mathrm{d}y + \int \epsilon_{\mathrm{xc}}[\rho(x)]\,\mathrm{d}x,\\
s.t. \quad \int \psi_i(x)\psi_j(x)\,\mathrm{d}x = \delta_{ij}, \quad i,j = 1,\ldots,N,\\
\rho(x) = \sum_{i=1}^{N} |\psi_i(x)|^2.
\end{aligned} \tag{1.19}$$

Here $\rho(x)$ is a function of $\{\psi_i\}$, and the minimization problem (1.19) is a nonlinear optimization problem. Eq. (1.19) can be solved directly using nonlinear optimization techniques [6, 25, 145, 207, 244, 250]. However, in practice it is more popular to solve the Euler-Lagrange equation associated to (1.19), which is called the self-consistent iteration. The self-consistent iteration method is used in this dissertation, and the basic procedure is summarized as follows.

The Euler-Lagrange equation corresponding to the minimization problem (1.19) is

$$\begin{aligned}
\left(-\frac{1}{2}\Delta + V_{\mathrm{eff}}[\rho]\right)\psi_i = \sum_{j=1}^{N} \psi_j \lambda_{ji},\\
s.t. \quad \int \psi_i(x)\psi_j(x)\,\mathrm{d}x = \delta_{ij}, \quad i,j = 1,\ldots,N,\\
\rho(x) = \sum_{i=1}^{N} |\psi_i(x)|^2.
\end{aligned} \tag{1.20}$$

We denote the effective potential by $V_{\mathrm{eff}}[\rho]$:

$$V_{\mathrm{eff}}[\rho](x) = V_{\mathrm{ext}}(x) + \int \frac{\rho(y)}{|x-y|}\,\mathrm{d}y + \epsilon'_{\mathrm{xc}}[\rho(x)]. \tag{1.21}$$

$\{\lambda_{ji}\}$ are the Lagrange multipliers corresponding to the orthonormal constraints of the electron orbitals. Eq. (1.20) is invariant under unitary transformations of the electron orbitals. As a result Eq. (1.20) can be simplified as

$$\begin{aligned} &\left(-\frac{1}{2}\Delta + V_{\text{eff}}[\rho]\right)\psi_i = \psi_i\epsilon_i, \\ s.t. \quad &\int \psi_i(x)\psi_j(x)\,\mathrm{d}x = \delta_{ij}, \quad i,j = 1,\ldots,N, \\ &\rho(x) = \sum_{i=1}^{N}|\psi_i(x)|^2. \end{aligned} \tag{1.22}$$

In order to minimize (1.19) only the lowest N eigenvalues and eigenvectors are to be computed. The lowest N eigenvalues $\{\epsilon_i\}_{i=1}^N$ are called the occupied Kohn-Sham eigenvalues, and the corresponding lowest N eigenvectors $\{\psi_i\}_{i=1}^N$ are called the occupied Kohn-Sham orbitals. The minimization problem (1.19) is nonlinear, and as a result the eigenvalue problem (1.22) is a nonlinear eigenvalue problem.

The Euler-Lagrange equation (1.22) can be solved by fixing the electron density $\rho = \rho_{\text{in}}$ in the potential energy term $V_{\text{eff}}[\rho]$. Then the Kohn-Sham Hamiltonian $H = -\frac{1}{2}\Delta + V_{\text{eff}}[\rho]$ is a fixed linear operator. The corresponding lowest N eigenvalues and eigenvectors can therefore be computed by a standard linear eigenvalue procedure such as ARPACK [151]. The consequence of fixing the electron density in the Kohn-Sham Hamiltonian is that the output electron density ρ_{out} given by Eq. (1.15) does not necessarily match the input electron density ρ_{in}. In such case, a new density ρ is generated based on ρ_{in} and ρ_{out}. This new density ρ is used as the new input density for the eigenvalue problem (1.22). This procedure is repeated until $\rho_{\text{in}} = \rho_{\text{out}}$. Since the self-consistent electron density is obtained iteratively, this procedure is called the self-consistent iteration.

When the self-consistent electron density ρ is obtained, the ground state electron energy can be calculated from the Kohn-Sham energies ϵ_i and the electron density ρ

according to:

$$E_{KS} = \sum_{i=1}^{N} \epsilon_i - \frac{1}{2} \iint \frac{\rho(x)\rho(y)}{|x-y|} \,\mathrm{d}x \,\mathrm{d}y + \int \epsilon_{\mathrm{xc}}[\rho(x)] \,\mathrm{d}x - \int \epsilon'_{\mathrm{xc}}[\rho(x)]\rho(x) \,\mathrm{d}x. \quad (1.23)$$

1.4 KSDFT: pseudopotential framework

So far KSDFT is introduced as an all electron theory. Namely all the electrons are taken into account in the calculation. In quantum chemistry, the electrons are divided into two categories: core electrons and valence electrons. For most of the systems, only the valence electrons participate in the interstitial bonding region and in the chemical reactions, and the core electrons do not participate in the chemical reactions. However, the electron orbitals of the core electrons are highly non-smooth and the treatment of the core electrons requires a large number of basis functions per atom or a fine mesh for numerical integration. Therefore it is desirable to remove the core electrons and represent the core electrons effectively in the potential energy surface. This idea is achieved in the *pseudopotential framework* [241, 245]. The pseudopotential framework only involves valence electrons. The number of basis functions per atom to describe the valence electrons is also much smaller than that in the all electron framework, and there is no singularity in the electronic wavefunctions. Pseudopotential framework will be used throughout this dissertation to describe the Kohn-Sham Hamiltonian unless otherwise specified. We remark that similar results can be achieved by projected augmented wavefunctions method (PAW) [34]. The extensions to other frameworks such as PAW and the all-electron framework will be the work in future.

In the past three decades, a vast number of types of pseudopotentials have been developed. The most widely used pseudopotentials include the norm-conserving pseudopotential [120, 241], the dual-space pseudopotential [104, 121] and the ultrasoft pseudopotential [245]. For a more detailed discussion and comparison of the pseudopotential theories, we refer the readers to the review article [59]. The Troullier-Martins pseudopotential [241] is one of the most popular norm-conserving pseudopotential. In what follows, the Troullier-Martins pseudopotential will be used for illustration purpose.

The Kohn-Sham energy functional in the Troullier-Martins pseudopotential framework is given by:

$$\begin{aligned} E_{\mathcal{K}}(\{\psi_i\}) = \frac{1}{2}\sum_{i=1}^{N}\int |\nabla\psi_i(x)|^2 \,\mathrm{d}x + \int V_{\text{ext}}(x)\rho(x)\,\mathrm{d}x + \sum_{\ell}\gamma_\ell\sum_{i=1}^{N}|\int b_\ell^*(x)\psi_i(x)\,\mathrm{d}x|^2 \\ + \frac{1}{2}\iint \frac{\rho(x)\rho(y)}{|x-y|}\,\mathrm{d}x\,\mathrm{d}y + \int \epsilon_{\text{xc}}[\rho(x)]\,\mathrm{d}x, \quad (1.24) \end{aligned}$$

In (1.24), we have taken the Kleinman-Bylander form of the pseudopotential [138]. For each ℓ, b_ℓ is a function supported locally in the real space around the position of one of the atoms, and $\gamma_\ell = \pm 1$.

The Kohn-Sham equation, or the Euler-Lagrange equation associated with (1.24) reads

$$H_{\text{eff}}[\rho]\psi_i = (-\tfrac{1}{2}\Delta + V_{\text{eff}}[\rho] + \sum_{\ell}\gamma_\ell|b_\ell\rangle\langle b_\ell|)\psi_i = \epsilon_i\psi_i, \quad (1.25)$$

where the effective one-body potential V_{eff} is given by

$$V_{\text{eff}}[\rho](x) = V_{\text{ext}}(x) + \int \frac{\rho(y)}{|x-y|}\,\mathrm{d}y + \epsilon'_{\text{xc}}[\rho(x)]. \quad (1.26)$$

After obtaining the self-consistent electron density, the total energy of the system can be expressed using the eigenvalues $\{\epsilon_i\}$ and ρ

$$E_{\text{KS}} = \sum_{i=1}^{N}\epsilon_i - \frac{1}{2}\iint \frac{\rho(x)\rho(y)}{|x-y|}\,\mathrm{d}x\,\mathrm{d}y + \int \epsilon_{\text{xc}}[\rho(x)]\,\mathrm{d}x - \int \epsilon'_{\text{xc}}[\rho(x)]\rho(x)\,\mathrm{d}x. \quad (1.27)$$

In each step of the self-consistent iteration, we find $\widetilde{\rho}$ from a given effective potential V_{eff}

$$\widetilde{\rho}(x) = \sum_{i=1}^{N} |\psi_i|^2(x), \tag{1.28}$$

where the $\{\psi_i\}$'s are the first N eigenfunctions of H_{eff}.

$$H_{\text{eff}}\psi_i = (-\tfrac{1}{2}\Delta + V_{\text{eff}} + \sum_{\ell} \gamma_\ell |b_\ell\rangle\langle b_\ell|)\psi_i = \epsilon_i \psi_i. \tag{1.29}$$

The $\{\psi_i\}$'s also minimize the variational problem

$$E_{\text{eff}}(\{\psi_i\}) = \frac{1}{2}\sum_{i=1}^{N} \int |\nabla\psi_i(x)|^2 \,\mathrm{d}x + \int V_{\text{eff}}(x)\rho(x)\,\mathrm{d}x + \sum_{\ell} \gamma_\ell \sum_{i=1}^{N} |\langle b_\ell, \psi_i\rangle|^2, \tag{1.30}$$

with the orthonormality constraints $\langle \psi_i, \psi_j \rangle = \delta_{ij}$.

1.5 Mathematical properties of KSDFT

In the self-consistent iteration framework for KSDFT, there are two major steps:

1. Given an input electron density ρ_{in}, calculate the output electron density ρ_{out}. This is done by solving the Kohn-Sham energies $\{\epsilon_i\}$ and the Kohn-Sham electron orbitals $\{\psi_i\}$ of $H[\rho_{\text{in}}]$.
2. Form a new input electron density from ρ_{in} and ρ_{out}.

In the standard method for solving KSDFT, the complexity of step 1 is $\mathcal{O}(N^3)$, and the complexity of step 2 is $\mathcal{O}(N)$. Therefore step 1 dominates the computational cost for solving KSDFT, and is the major bottleneck in order to reduce the complexity.

Step 1 essentially defines a map from ρ_{in} to ρ_{out}, which is referred to as the *Kohn-Sham map*. Step 1 involves a diagonalization process of $H[\rho_{\text{in}}]$ which is a nonlinear process. Therefore the Kohn-Sham map is a nonlinear map. Kohn-Sham map contains all the information of step 1. The mathematical properties of the

Kohn-Sham map are essential in order to achieve an accurate and efficient method for solving KSDFT.

In order to study the mathematical properties of the Kohn-Sham map, it is desirable to have the explicit form of the Kohn-Sham map, rather than the implicit form as defined in step 1. The explicit form of the Kohn-Sham map is as follows. For simplicity we assume the temperature is zero. The Hamiltonian matrix is denoted by $H \equiv H[\rho_{\mathrm{in}}]$ which is discretized into a $N_t \times N_t$ matrix. $\{\epsilon_i\}_{i=1}^{N_t}$ and $\{\psi_i\}_{i=1}^{N_t}$ are all the eigenvalues and eigenvectors of the Hamiltonian matrix H. The output electron density can be rewritten in an alternative form:

$$\begin{aligned}\rho_{\mathrm{out}}(x) &= \sum_{i=1}^{N} |\psi_i(x)|^2 \\ &= \begin{pmatrix} \psi_1(x) & \cdots & \psi_{N_t}(x) \end{pmatrix} \begin{pmatrix} \chi(\varepsilon_1 - \mu) & & \\ & \ddots & \\ & & \chi(\varepsilon_{N_t} - \mu) \end{pmatrix} \begin{pmatrix} \psi_1(x) \\ \vdots \\ \psi_{N_t}(x) \end{pmatrix}\end{aligned} \tag{1.31}$$

Here $\chi(x)$ is the Heaviside function that satisfies

$$\chi(x) = \begin{cases} 1, & x \le 0, \\ 0, & x > 0. \end{cases} \tag{1.32}$$

μ is called the chemical potential. For a discretized system, μ is chosen to be in the range $(\epsilon_N, \epsilon_{N+1})$ as long as $\epsilon_{N+1} > \epsilon_N$. Eq. (1.31) can be written in a more compact form using the notation of matrix function:

$$\rho_{\mathrm{out}}(x) = \left[\chi\left(H[\rho_{\mathrm{in}}] - \mu I\right)\right]_{x,x} \equiv \mathrm{diag}\ \chi(H[\rho_{\mathrm{in}}] - \mu I). \tag{1.33}$$

Here χ is a matrix function and I is the identity matrix of size $N_t \times N_t$. Eq. (1.33) clearly shows that the Kohn-Sham map is nothing but the diagonal elements of the matrix Heaviside function $\chi(H[\rho_{\mathrm{in}}] - \mu I)$.

However, the value of the Heaviside function χ is either 0 or 1 on the spectrum of the Hamiltonian matrix. The Heaviside function is a not a smooth function, and the matrix Heaviside function is not well-defined for all systems. Here the important characteristic quantity is $\epsilon_{N+1} - \epsilon_N$, which is referred to as the *energy gap* of the system. It can be shown that as the number of electrons $N \to \infty$, the energy gap is always finite for insulating systems, and becomes zero for metallic systems [12]. As a result, the matrix Heaviside function is only well defined for insulating system, and is ill-defined for metallic systems.

The flaw of the matrix Heaviside function can be amended by a more generalized function called the matrix Fermi-Dirac function, which takes into account the finite temperature effect [186]

$$\rho = \operatorname{diag} \ \frac{1}{1 + \exp(\beta(H - \mu))} \equiv \operatorname{diag} \ f(H) \tag{1.34}$$

Fermi-Dirac function is closely related to the Heaviside function: If β is finite, the Fermi-Dirac function is a smooth function across the spectrum of the Hamiltonian H, and is well-defined regardless of the value of the energy gap. When $\beta \to \infty$, Fermi-Dirac function converges to the Heaviside function (see Fig. 1.1). The physical meaning of β is the inverse of the temperature of the system, and $\beta \to \infty$ implies that the temperature is zero. Therefore the matrix Heaviside function is also called the *zero temperature limit of the Fermi-Dirac function.*

The ground state energy E_{KS} can be written in terms of $f(H)$ as well. For insulating systems, we have

$$\sum_{i=1}^{N} \epsilon_i = \operatorname{Tr}\left[H\chi(H - \mu I)\right], \tag{1.35}$$

and this relation can be directly generalized to both insulating systems and metallic systems as $\operatorname{Tr}\left[Hf(H)\right]$. Thus, the matrix function $f(H)$ is of central importance

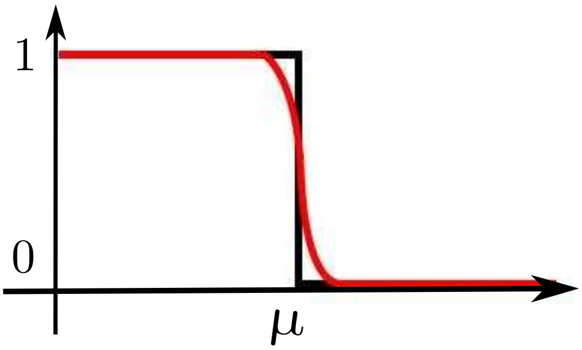

Figure 1.1: Heaviside function (black line) and Fermi-Dirac function at finite temperature (red line).

in KSDFT, and is referred to as the *density matrix* of the system. KSDFT can be written explicitly only using density matrix:

$$\begin{aligned}\rho =& \operatorname{diag}\; f(H[\rho]), \qquad (1.36)\\ E_{KS} =& \operatorname{Tr}\left[H[\rho]f(H[\rho])\right] - \frac{1}{2}\iint \frac{\rho(x)\rho(y)}{|x-y|}\,\mathrm{d}x\,\mathrm{d}y\\ &+ \int \epsilon_{\mathrm{xc}}[\rho(x)]\,\mathrm{d}x - \int \epsilon'_{\mathrm{xc}}[\rho(x)]\rho(x)\,\mathrm{d}x. \qquad (1.37)\end{aligned}$$

The density matrix $f(H)$ is a $N_t \times N_t$ matrix. Eq. (1.37) shows the remarkable property that not all the elements of the density matrix are required in KSDFT. Electron density ρ only requires the diagonal elements of the density matrix. The Hamiltonian matrix H contains the Laplacian operator, and the nearest off-diagonal elements of the density matrix are needed to calculate $\operatorname{Tr}[Hf(H)]$ in the ground state energy. In summary, KSDFT *only requires the diagonal elements and the nearest off-diagonal elements of the density matrix.* This property is essential in order to achieve an accurate and efficient method for solving KSDFT.

Although this mathematical property has been observed for a long time [101], it is not at all reflected in the existing methods for KSDFT calculation. The direct reason is that it is not straightforward to calculate the diagonal and nearest off-diagonal elements of a complicated matrix function. Before introducing the new method developed in this dissertation that directly calculates the diagonal and nearest off-diagonal elements of the density matrix, we first discuss the existing methods for solving KSDFT.

1.6 Existing methods and software packages for solving KSDFT

Most of the existing methods for solving KSDFT can be categorized into two types: cubic scaling methods and linear scaling methods. Within each category there are a large number of software packages available. In this section we mainly discuss two types of algorithms in the pseudopotential framework. To facilitate readers who are unfamiliar with this subject, a short list of the capabilities for the most versatile software packages is given at the end of the discussion. The URL address of each software package is provided for readers who are interested in further details. Finally, for completeness of the discussion we also mention some software packages for all-electron calculations.

1.6.1 Cubic scaling methods

Cubic scaling method is implemented in most of the popular software packages for KSDFT calculation. The cubic scaling methods include the direct diagonalization methods such as the Davidson method [69], the conjugate gradient method (CG) [239], and the direct inversion in the iterative subspace (DIIS) method [214]. Other variants that also fall into this category include the Car-Parrinello method [49] and the Chebyshev filtering method [256], to name a few. The variants of the diagonalization methods result in different preconstant in front of the asymptotic cubic scaling. However, the orthogonalization step is inevitable in order to obtain the electron density and the ground state energy, and leaves the cubic scaling unchanged.

Take the diagonalization method for instance, the lowest $\widetilde{N}$ eigenvalues where $\widetilde{N} = N$ for insulating system or $\widetilde{N} > N$ for metallic system (to include the finite temperature effect) are directly computed. The Kohn-Sham map is then evaluated as

$$\rho(x) = \sum_{i=1}^{\widetilde{N}} \frac{1}{1 + e^{\beta(\epsilon_i - \mu)}} |\psi_i(x)|^2. \tag{1.38}$$

Since the diagonalization method constructs all the occupied electron orbitals $\{\psi_i\}$ explicitly, the entire density matrix $f(H)$ is essentially constructed. It does not take advantage of the mathematical property that only the diagonal elements and nearest off-diagonal elements of the density matrix are needed in KSDFT calculation. The diagonalization method involves an orthogonalization step of the subspace $\Psi = [\psi_1, \cdots, \psi_{\widetilde{N}}]$. In the discrete case, the length of each vector ψ_i is proportional to N and the total number of electrons is N. Therefore the orthogonalization step scales as $\mathcal{O}(N^3)$ with respect to the number of electrons in the system, and the computational cost of KSDFT becomes very high for large number of electrons.

Below are some representative software packages for electronic structure calculation using cubic scaling methods:

- ABINIT: Diagonalization method with planewave basis functions.

  `http://www.abinit.org/`

- BigDFT: Diagonalization method with a two-level wavelet basis functions.

  `http://inac.cea.fr/L\_Sim/BigDFT/`

- CASTEP: Diagonalization method with planewave basis functions.

  `http://www.castep.org/`

- CP2K: Diagonalization method with mixed Gaussian and planewave basis functions.

`http://cp2k.berlios.de/`

- CPMD: Diagonalization method as well as Car-Parrinello method with planewave basis functions.

  `http://www.cpmd.org/`

- OPENMX (Open source package for material explorer): Diagonalization method with planewave basis functions and numerical atomic orbitals.

  `http://www.openmx-square.org/`

- PASRSEC (Pseudopotential Algorithm for Real-Space Electronic Calculations): Diagonalization method and Chebyshev filtering method with finite difference discretization.

  `http://parsec.ices.utexas.edu/index.html`

- Quantum ESPRESSO: Diagonalization method as well as Car-Parrinello method with planewave basis functions.

  `http://www.quantum-espresso.org/`

- VASP (Vienna Ab-initio Simulation Package): Diagonalization method with planewave basis functions.

  `http://cms.mpi.univie.ac.at/vasp/`

1.6.2 Linear scaling methods

The major breakthrough that reduces the $\mathcal{O}(N^3)$ in the past two decades is the linear scaling methods. The linear scaling methods use the nearsightedness property, which means that the density perturbation induced by a local change in the external potential decays off exponentially with respect to the distance from the place where the perturbation was applied, and also that the off-diagonal elements of the density matrix decay exponentially [142, 212]. The nearsightedness property is valid for insulating systems and metallic systems at finite temperature. The nearsightedness property is not valid for metallic systems at zero temperature due to the well-known Friedel oscillation [87]. Due to the fast decay of the density matrix along the off-diagonal direction, the density matrix can be truncated beyond a certain range along the off-diagonal direction for insulating systems. Various methods have been proposed based on different perspectives of the nearsightedness property (for a detailed review, see [101]). We review briefly some representative linear scaling methods as below. The linear scaling methods are mainly divided into two classes.

The first class of linear scaling algorithms are based on the localization of electron orbitals and subspaces of electron orbitals. In the orbital minimization approach (OM) [88, 184], the truncation of the electron orbitals is imposed by adding an additional confining potential to the Hamiltonian. Orbital minimization approach can have multiple minima [136]. The orbital minimization method can be combined with the localization procedure (OML) [89] to eliminate the multiple minima problem. The localized subspace iteration method (LSI) [91] localizes the subspace consisting several electron orbitals, and obtains the optimal truncation radius.

The second class of linear scaling algorithms are based on the localization of the density matrix directly. In the divide-and-conquer method (D&C) [251], the electron density is divided into a set of loosely coupled subsystems. Each subsystem is solved separately by standard diagonalization methods and linear scaling is achieved. The density matrix minimization method (DMM) [154, 229] achieves linear scaling by directly truncating the density matrix beyond a predetermined truncation radius, with the help of the McWeeny purification transformation [185]. The density matrix is then optimized using nonlinear conjugate gradient method. The Fermi operator expansion method (FOE) [19, 99] expands the Fermi-Dirac matrix function into simple matrix functions that can be directly evaluated without diagonalization of the

Hamiltonian. These simple matrix functions can be polynomials or rational functions of the Hamiltonian matrix. Each simple matrix function is only evaluated within the truncation range of the density matrix along the off-diagonal direction, and the FOE method achieves linear scaling.

Several widely used linear scaling methods for the electronic structure calculation of insulating systems include:

- CONQUEST: b-spline basis functions and Pseudo-atomic orbitals (PAO). Linear scaling is achieved by McWeeny's purification method [185].

  `http://hamlin.phys.ucl.ac.uk/NewCQWeb/bin/view`

- ONETEP (Order-N Electronic Total Energy Package): Non-orthogonal generalized Wannier functions (NGWF). Linear scaling is achieved by density kernel optimization method which is a variant of the density matrix minimization method [154].

  `http://www2.tcm.phy.cam.ac.uk/onetep/`

- SIESTA (Spanish Initiative for Electronic Simulations with Thousands of Atoms): Numerical atomic orbitals. Linear scaling is achieved by orbital minimization method [184].

  `http://www.icmab.es/siesta/`

1.6.3 All-electron methods

As opposed to the pseudopotential framework which only involves valence electrons, all electron methods treat the core electrons and the valence electrons on the same footing. All electron methods can avoid the numerical error caused by the limited transferability in the pseudopotentials, but the computational cost of the all-electron methods is generally significantly larger than that of the pseudopotential methods.

The increased computational cost is mainly due to the fact that the electron orbitals are non-smooth at the positions of the nuclei (satisfying the cusp condition [134]) and are much more oscillatory around the positions of the nuclei. As a result, analytic or semi-analytic forms of basis functions are generally used in the all-electron calculations, such as Slater-type orbitals (STO), Gaussian-type orbitals (GTO) and numerical atomic orbitals (NAO). Several widely used software packages for all-electron calculation include:

- ADF (Amsterdam Density Functional): Diagonalization methods with Slater-type orbitals (STO).

  `http://www.scm.com/`

- GAUSSIAN: Diagonalization methods with Gaussian-type orbitals (GTO).

  `http://www.gaussian.com/`

- FHI-aims: Diagonalization methods with numerical atomic orbitals (NAO).

  `https://aimsclub.fhi-berlin.mpg.de/`

- WIEN2k: Diagonalization methods with full-potential (linearized) augmented plane-wave (FP-LAPW) plus local orbitals (LO) basis functions.

  `http://www.wien2k.at/`

1.7 Unified, accurate and efficient method for solving KSDFT

Section 1.6 summarizes the most widely used numerical schemes for solving KSDFT. Cubic scaling methods involve the orthogonalization step which is intrinsically of $\mathcal{O}(N^3)$ scaling and is difficult to be improved in general. Linear scaling methods uses the nearsightedness property for insulating systems and metallic systems at finite temperature. However, the performance of the linear scaling methods relies crucially on the truncation radius. The truncation radius can be quite large if high numerical accuracy is to be achieved, especially for metallic system where the truncation radius depends explicitly on the artificial temperature. Therefore linear scaling algorithms typically exhibit advantage only for systems with a very large number of electrons [101]. Moreover, linear scaling methods still construct the entire density matrix, and they do not aim at calculating the diagonal elements and nearest off-diagonal elements of the density matrix directly. In order to avoid the difficulties in the linear scaling methods, it is desirable to design a method that does not use nearsightedness, and that calculates the diagonal elements and nearest off-diagonal elements of the density matrix directly. Part I of this dissertation achieves this goal under the framework of Fermi operator expansion (FOE).

FOE expands the Fermi-Dirac matrix function into simple matrix functions. Each simple matrix function is calculated directly without diagonalization process, and thus FOE does not involve the orthogonalization step. In calculating simple matrix functions, FOE does not necessarily require the nearsightedness property. As a result FOE has none of the previously mentioned drawbacks. The new method is accurate, efficient, and is applicable to both insulating and metallic systems at low or at high temperature. First we discuss the basic procedure of FOE.

FOE expands the Fermi-Dirac matrix function $f(H)$ into simple matrix functions $\{f_i(H)\}$, *i.e.*

$$f(H) \approx \sum_{i=1}^{P} f_i(H), \tag{1.39}$$

and the electron density and the ground state energy can be calculated as

$$\rho \approx \sum_{i=1}^{P} \operatorname{diag}\; f_i(H[\rho]), \tag{1.40}$$

$$\begin{aligned} E_{KS} \approx & \sum_{i=1}^{P} \operatorname{Tr}\left[H[\rho] f_i(H[\rho])\right] - \frac{1}{2} \iint \frac{\rho(x)\rho(y)}{|x-y|} \,\mathrm{d}x\,\mathrm{d}y \\ & + \int \epsilon_{\mathrm{xc}}[\rho(x)]\,\mathrm{d}x - \int \epsilon'_{\mathrm{xc}}[\rho(x)]\rho(x)\,\mathrm{d}x. \end{aligned} \tag{1.41}$$

Therefore under the framework of FOE, only the diagonal elements and nearest off-diagonal elements of each matrix function $f_i(H)$ are to be calculated. The calculation of $f_i(H)$ for different i are independent from each other.

Each simple matrix function $f_i(H)$ should take certain form in order to be evaluated directly without diagonalization. To be more specific, $f_i(H)$ can only be polynomial matrix function or rational matrix function. FOE based on both the polynomial expansion [16, 99, 102, 155, 156] and the rational expansion [19, 103, 144, 160, 164, 199, 227] have been developed. FOE based on the polynomial expansion requires calculating the powers of H. FOE based on the rational expansion requires calculating the inverse of H. Both operations are $\mathcal{O}(N^3)$ without further simplification, and FOE does not exhibit advantage over diagonalization methods for metallic systems. As a result, FOE is only mentioned sporadically in literature for certain classes of systems [144, 227].

Part I of this dissertation develops FOE to be an accurate and efficient method for solving KSDFT in all systems. The new method achieves uniform improvement over the $\mathcal{O}(N^3)$ method for systems under all dimensions. The asymptotic scaling of the new method is $\mathcal{O}(N)$ for one dimensional system, $\mathcal{O}(N^{1.5})$ for two-dimensional system, and $\mathcal{O}(N^2)$ for three-dimensional system. Furthermore, the new method can be combined with the nearsightedness property to achieve $\mathcal{O}(N)$ scaling at all dimensions for insulating systems and for metallic systems at high temperature.

We also expect that the new method should not only exhibit improved asymptotic scaling but also have a relatively small preconstant. To this end it is necessary to systematically study all the phases of FOE. In this dissertation, the complete flowchart of FOE is divided into four phases (see Fig. 1.2): Discretize the Hamiltonian operator H into a matrix of finite size (discretization); Represent the Fermi-Dirac matrix function $f(H)$ into appropriate simple matrix functions $\{f_i(H)\}$ (representation); Evaluate the diagonal and nearest off-diagonal elements of each $\{f_i(H)\}$ (evaluation); Self-consistent iteration (iteration). Part I of this dissertation develops a novel scheme, named the adaptive local basis functions for the discretization step. The adaptive local basis functions achieve high accuracy (below 10^{-3} Hartree/atom) with a very small number of basis functions. This dissertation presents the optimal strategy for the representation step, which represents the Fermi-Dirac operator in terms of a simple rational expansion called the pole expansion. This dissertation further develops a fast algorithm for evaluating the diagonal and nearest off-diagonal elements of each rational function, called the selected inversion algorithm. The computational scaling of the selected inversion algorithm to evaluate each rational function is $\mathcal{O}(N)$ for one dimensional systems, $\mathcal{O}(N^{1.5})$ for two-dimensional systems, and $\mathcal{O}(N^2)$ for three-dimensional systems. Self-consistent iteration is an important component in the KSDFT calculation. However, the self-consistent iteration does not cause the $\mathcal{O}(N^3)$ scaling problem and is a relatively separate issue. The self-consistent iteration is not discussed in this dissertation, but will be studied in the future work.

The rest of Part I of this dissertation is organized as follows. Chapter 2 discusses the discretization technique for KSDFT, and introduces the novel adaptive local basis functions. Chapter 3 discusses various representation methods of the Fermi-Dirac operator, and presents the optimal strategy for representing the Fermi-Dirac operator in terms of rational expansion. Chapter 4 introduces a new methodology named selective inversion for evaluating the diagonal elements and nearest off-diagonal ele-

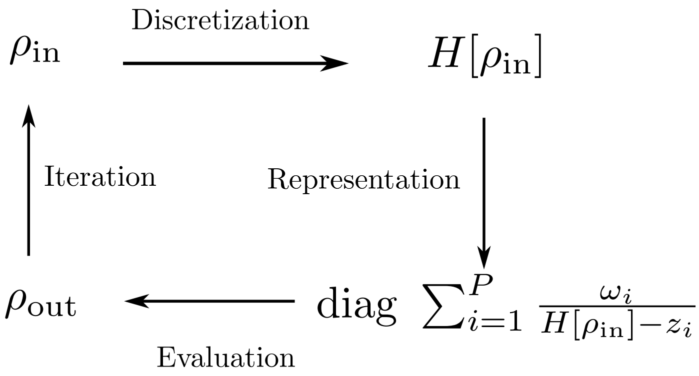

Figure 1.2: Flowchart of the unified, accurate and efficient method developed in this dissertation for solving KSDFT.

ments of each simple matrix function. The work of selected inversion also inspired us developing a fast algorithm for the construction of hierarchical matrices [161]. This is illustrated in Chapter 5. Finally, Chapter 6 concludes Part I of the dissertation with open questions and further work.