SIAM J. NUMER. ANAL.
Vol. 53, No. 4, pp. 1758–1785

ON THE ANALYSIS OF THE DISCRETIZED KOHN–SHAM DENSITY FUNCTIONAL THEORY*

XIN LIU†, ZAIWEN WEN‡, XIAO WANG§, MICHAEL ULBRICH¶, AND YAXIANG YUAN†

Abstract. In this paper, we study a few theoretical issues in the discretized Kohn–Sham (KS) density functional theory. The equivalence between either a local or global minimizer of the KS total energy minimization problem and the solution to the KS equation is established under certain assumptions. The nonzero charge densities of a strong local minimizer are shown to be bounded from below by a positive constant uniformly. We analyze the self-consistent field (SCF) iteration by formulating the KS equation as a fixed point map with respect to the potential. The Jacobian of these fixed point maps is derived explicitly. Both global and local convergence of the simple mixing scheme can be established if the gap between the occupied states and unoccupied states is sufficiently large. This assumption can be relaxed in certain cases. Numerical experiments based on the MATLAB toolbox KSSOLV show that it holds on a few simple examples. Although our assumption on the gap is very stringent, our analysis is still valuable for a better understanding of the KS minimization problem, the KS equation, and the SCF iteration.

Key words. Kohn–Sham total energy minimization, Kohn–Sham equation, self-consistent field iteration, nonlinear eigenvalue problem

AMS subject classifications. 15A18, 65F15, 47J10, 90C30

DOI. 10.1137/140957962

1. Introduction.

The Kohn–Sham (KS) density functional theory in electronic structure calculations can be formulated as either a total energy minimization problem or a nonlinear eigenvalue problem. Using a suitable discretization scheme whose spatial degree of freedom is n, the electron wave functions of p occupied states can be approximated by a matrix $X = [x_1, \ldots, x_p] \in \mathbb{R}^{n\times p}$. The charge density of electrons associated with the occupied states is defined as

$$\rho(X) := \mathrm{diag}(XX^{\mathrm{T}}), \tag{1.1}$$

where $\mathrm{diag}(A)$ denotes the vector containing the diagonal elements of the matrix A. Let $\mathrm{tr}(A)$ be the trace of $A \in \mathbb{R}^{n\times n}$, i.e., the sum of the diagonal elements of A. A commonly used discretized KS total energy function has the form

$$E(X) := \frac{1}{4}\mathrm{tr}(X^{\mathrm{T}}LX) + \frac{1}{2}\mathrm{tr}(X^{\mathrm{T}}V_{ion}X) + \frac{1}{4}\rho^{\top}L^{\dagger}\rho + \frac{1}{2}e^{\mathrm{T}}\epsilon_{xc}(\rho), \tag{1.2}$$

*Received by the editors February 21, 2014; accepted for publication (in revised form) May 22, 2015; published electronically July 14, 2015.
http://www.siam.org/journals/sinum/53-4/95796.html

†State Key Laboratory of Scientific and Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China (liuxin@lsec.cc.ac.cn, yyx@lsec.cc.ac.cn). The research of the first author was supported in part by NSFC grants 11101409 and 11331012, and the National Center for Mathematics and Interdisciplinary Sciences, CAS. The research of the fifth author was supported in part by NSFC grant 11331012.

‡Beijing International Center for Mathematical Research, Peking University, China (wenzw@math.pku.edu.cn). The research of this author was supported in part by NSFC grants 11322109 and 91330202 and by the National Basic Research Project under the grant 2015CB856000.

§School of Mathematical Sciences, University of Chinese Academy of Sciences, China (wangxiao@ucas.ac.cn). The research of this author was supported in part by Postdoc grant 119103S175, UCAS president grant Y35101AY00, and NSFC grant 11301505.

¶Chair of Mathematical Optimization, Department of Mathematics, Technische Universität München, 85747 Garching b. München, Germany (mulbrich@ma.tum.de).

where L is a finite-dimensional representation of the Laplacian operator, V_{ion} is the ionic pseudopotentials sampled on a suitably chosen Cartesian grid, $L^\dagger$ corresponds to the pseudoinverse of L, e is the column vector of all ones, and $\epsilon_{xc}(\rho) \in \mathbb{R}^n$ denotes the exchange correlation energy functional. The four terms in $E(X)$ describe the kinetic energy, local ionic potential energy, Hartree potential energy, and exchange correlation energy, respectively.

The KS total energy minimization problem solves

$$\begin{aligned} \min_{X \in \mathbb{R}^{n\times p}} \quad & E(X) \\ \text{s.t.} \quad & X^{\mathrm{T}}X = I. \end{aligned} \tag{1.3}$$

The orthogonality constraints are imposed since the wave functions X must be orthogonal to each other due to physical constraints. It can be verified that the gradient of $E(X)$ with respect to X is $\nabla E(X) = H(X)X$, where the Hamiltonian $H(X) \in \mathbb{R}^{n\times n}$ is a matrix function

$$H(X) := \frac{1}{2}L + V_{ion} + \mathrm{Diag}(L^\dagger \rho) + \mathrm{Diag}(\mu_{xc}(\rho)^{\mathrm{T}} e), \tag{1.4}$$

where $\mu_{xc}(\rho)$ is the Jacobian of ϵ_{xc} with respect to ρ, i.e.,

$$\mu_{xc}(\rho) = \left(\frac{\partial \epsilon_{xc}}{\partial \rho}\right)^{\mathrm{T}} = \left[\frac{\partial(\epsilon_{xc})_1}{\partial \rho}, \ldots, \frac{\partial(\epsilon_{xc})_n}{\partial \rho}\right]^{\mathrm{T}} \in \mathbb{R}^{n\times n},$$

and $\mathrm{Diag}(x)$ denotes a diagonal matrix with x on its diagonal. The so-called KS equation is

$$\begin{aligned} H(X)X &= X\Lambda, \\ X^{\mathrm{T}}X &= I, \end{aligned} \tag{1.5}$$

where Λ is a diagonal matrix consisting of p smallest eigenvalues of $H(X)$ if there is a gap between the pth and $(p+1)$st smallest eigenvalues of $H(X)$. The definition of the KS equation is slightly different if this gap is zero; see the Fermi–Dirac distribution in subsection 3.1. The KS equation (1.5) is closely related to the first-order optimality conditions for (1.3) which are the same as (1.5) except that the diagonal matrix Λ consists of any p eigenvalues of $H(X)$.

In this paper, we first study the relationship between the KS total energy minimization problem (1.3) and the KS equation (1.5) under certain conditions. A simple counterexample is provided to demonstrate that the solutions of these two problems are not necessarily the same. The second-order optimality conditions for (1.3) are examined based on the assumption of the existence of the second-order derivative of the exchange correlation functional [17, 30]. For a specialized exchange correlation functional, we prove that a global solution of (1.3) is a solution of (1.5) if the gap between the pth and $(p+1)$st smallest eigenvalues of the Hamiltonian $H(X)$ is sufficiently large. The equivalence between a local minimizer of (1.3) and the solution (1.5) needs an additional assumption that the corresponding charge densities are all positive. For a strong local minimizer X^* which is defined based on the second-order sufficient optimality conditions for (1.3), we show that the nonzero charge densities at X^* are bounded from below by a positive constant uniformly.

Our second purpose is the analysis of the most widely used approach, the self-consistent field (SCF) iteration, for solving the KS equation (1.5). The SCF iteration

is based on computing a sequence of linear eigenvalue problems iteratively. It is well known that the basic version of the SCF iteration often converges slowly or fails to converge [19] even with the help of various heuristics. A convergence analysis of the SCF iteration for solving the Hartree–Fock equations according to the optimal damping algorithm (ODA) is established in [7] and an analysis of gradient-based algorithms for the Hartree–Fock equations is proposed in [22] using the Lojasiewiscz inequality. The interested reader is referred to [3, 4, 5, 6, 8, 9, 10, 11, 13, 14, 21, 27] for discussions on ODA, the gradient-based algorithms, and numerical analysis of density functional theory (DFT). A condition is identified in [32] such that the SCF iteration is a contractive fixed point iteration under a specific form of the Hamiltonian without involving any exchange correlation term. Global and local convergence of the SCF iteration for general KS DFT is established in [25] from an optimization point of view. Their assumptions include that the second-order derivative of the exchange correlation energy functional is uniformly bounded from above and the gap between the pth and $(p+1)$st eigenvalues of the Hamiltonian $H(X)$ is sufficiently large. A detailed overview is presented in Table 3 in the conclusion.

We improve the convergence results of the SCF iteration from the following three perspectives. (i) The KS equation (1.5) is formulated as a nonlinear system of equations (fixed point maps) with respect to either the charge density or potential. Applying the differentiability of spectral operators, the Jacobian (the "Adler–Wiser formula" [1, 31]) of these fixed point maps is derived explicitly and analyzed. (ii) Global convergence (i.e., convergence to a stationary point from any initial solution) of the simple mixing scheme can be established when there exists a gap between pth and $(p+1)$st eigenvalues of the Hamiltonian $H(X)$. This assumption can be relaxed for local convergence analysis, i.e., convergence behavior if the initial point is selected in a neighborhood sufficiently close to the solution of (1.5). If the charge density is computed using the Fermi–Dirac distribution, the assumption on the gap is not needed as long as a suitable step size for simple mixing is chosen. Our results require much weaker conditions than the previous analysis in [25]. (iii) We propose two approximate Newton methods according to the structure of the Jacobian of the fixed point maps and preliminary convergence results are also established for them. The second type of our approaches is exactly the method of Kerker preconditioner discussed in [24]. Although our assumption on the gap between eigenvalues of the Hamiltonian is very stringent, numerical experiments in the MATLAB toolbox KSSOLV [33] show that it holds on three simple examples, "nic," "si2h4," and "sih4." Hence, our analysis is still valuable for a better understanding of the KS equation and the SCF iteration.

The rest of this paper is organized as follows. A counterexample between the equivalence of the KS minimization and KS equation is presented in subsection 2.1. The optimality conditions for the KS minimization problem under smoothness assumptions on the exchange functional are provided in subsection 2.2. A condition for the equivalence between a local minimizer of the KS minimization and the KS equation is established in subsection 2.3. The corresponding analysis for a global minimizer is established in subsection 2.4. Lower bounds for the charge density at local minimizers are presented in subsection 2.5. In subsection 3.1, we view the KS equation as fixed point maps with respect to the charge density or potential. The Jacobian of these fixed point maps is presented in subsection 3.2. In section 4, we establish both local and global convergence for the SCF iteration with simple mixing schemes. Two approximate Newton approaches and their convergence properties are discussed in section 5. Finally, our assumptions on the eigenvalue gap are verified numerically on a few simple examples in section 6.

2. Equivalence between the KS total energy minimization and the KS equation.

2.1. A counterexample. The following three-dimensional (3D) toy example shows that a solution of the KS equation is not necessary for a global optimal solution of the KS total energy minimization problem. Let $n = 3$, $p = 1$, and choose

$$L = \begin{pmatrix} 1.4299 & -0.2839 & -0.4056 \\ -0.2839 & 1.1874 & 0.2678 \\ -0.4056 & 0.2678 & 1.3826 \end{pmatrix}, \quad V_{ion} = 0, \text{ and } \epsilon_{xc}(\rho) = 0.$$

It can be verified numerically that $X^* = \begin{pmatrix} 0.3683 & -0.6188 & 0.6939 \end{pmatrix}^{\mathrm{T}}$ is a global minimizer of (1.3). On the other hand, we have

$$H(X^*) = \frac{1}{2}L + \mathrm{Diag}(L^\dagger \rho(X^*)) = \begin{pmatrix} 0.9735 & -0.1419 & -0.2028 \\ -0.1419 & 0.8955 & 0.1339 \\ -0.2028 & 0.1339 & 1.0569 \end{pmatrix},$$

and X^* is an eigenvector associated with the second smallest eigenvalue of $H(X^*)$. Therefore, the equivalence between the KS total energy minimization and the KS equation only holds under certain assumptions. For this counterexample, our assumptions in subsections 2.3 and 2.4 do not hold because the gap between the the first and second eigenvalues of $H(X^*)$ is $\delta = 0.046$ and it is smaller than $||L^\dagger||_2 = 1$. We should point out that the above example may not exist in the practice of DFT.

2.2. Optimality conditions under smoothness assumptions on $\epsilon_{xc}(\rho)$. The Lagrangian function of the minimization problem (1.3) is $\mathcal{L}(X, \Lambda) := E(X) - \frac{1}{2}\mathrm{tr}(\Lambda(X^{\mathrm{T}}X - I))$. Suppose X is a local minimizer of (1.3). It follows from $X^{\mathrm{T}}X = I$ that the linear independence constraint qualification is satisfied. Hence, there exists a Lagrange multiplier Λ such that the first-order optimality conditions hold:

$$\nabla_X \mathcal{L}(X, \Lambda) = H(X)X - X\Lambda = 0 \text{ and } X^{\mathrm{T}}X = I. \tag{2.1}$$

Multiplying both sides of the first equality in (2.1) by X^{T} and using $X^{\mathrm{T}}X = I$, we have $\Lambda = X^{\mathrm{T}}H(X)X$, which is a symmetric matrix. Note that $E(XQ) = E(X)$ and $H(XQ) = H(X)$ hold for any orthogonal matrix $Q \in \mathbb{R}^{p\times p}$. Hence, if X is a stationary point, any matrix in the set $\{XQ \mid Q \in \mathbb{R}^{p\times p} \text{ and } Q^{\mathrm{T}}Q = I\}$ is also a stationary point, and their objective values are the same. Let $\tilde{Q}\tilde{\Lambda}\tilde{Q}^{\mathrm{T}}$ be the eigenvalue decomposition of $X^{\mathrm{T}}H(X)X$ and $\tilde{X} := X\tilde{Q}$. Then the Lagrangian multiplier $\tilde{\Lambda} = \tilde{X}^{\mathrm{T}}H(\tilde{X})\tilde{X}$ is a diagonal matrix whose entries are the eigenvalues of $H(X)$.

Let $\mathcal{L}(\mathbb{R}^{n\times p}, \mathbb{R}^{n\times p})$ denote the space of linear operators which map $\mathbb{R}^{n\times p}$ to $\mathbb{R}^{n\times p}$. The Fréchet derivative of $\nabla E(X)$ is defined as the (unique) function $\nabla^2 E : \mathbb{R}^{n\times p} \to \mathcal{L}(\mathbb{R}^{n\times p}, \mathbb{R}^{n\times p})$ such that

$$\lim_{\|S\|_{\mathrm{F}}\to 0} \frac{\|\nabla E(X + S) - \nabla E(X) - \nabla^2 E(X)(S)\|_{\mathrm{F}}}{\|S\|_{\mathrm{F}}} = 0.$$

The next lemma shows an explicit form of the Hessian operator, if the exchange correlation energy is second-order differentiable.

Lemma 2.1 (Lemma 2.1 in [30]). *Suppose that $\epsilon_{xc}(\rho(X))$ is twice differentiable with respect to $\rho(X)$. Let $\frac{\partial \mu_{xc}(\rho)}{\partial \rho} e$ be defined as $\sum_{i=1}^{n} \frac{\partial^2 (\epsilon_{xc})_i}{\partial \rho^2}$. Given a direction $S \in \mathbb{R}^{n\times p}$, the Hessian-vector product of $E(X)$ is*

$$\nabla^2 E(X)[S] = H(X)S + 2\mathrm{Diag}\left(J(\rho)\,\mathrm{diag}(SX^{\mathrm{T}})\right)X, \tag{2.2}$$

where

$$J(\rho) := L^\dagger + \frac{\partial \mu_{xc}(\rho)}{\partial \rho} e. \tag{2.3}$$

Consequently, the second-order necessary and sufficient optimality conditions can be obtained from Theorems 12.5 and 12.6 in [26], respectively.

THEOREM 2.2.

(1) *Suppose that X is a local minimizer of problem* (1.3) *and $\epsilon_{xc}(\rho(X))$ is twice differentiable with respect to $\rho(X)$. Then, for all $S \in \mathcal{T}(X)$, it holds*

$$\operatorname{tr}(S^{\mathrm{T}} H(X) S - \Lambda S^{\mathrm{T}} S) + 2 \operatorname{diag}(XS^{\mathrm{T}})^{\mathrm{T}} J \operatorname{diag}(XS^{\mathrm{T}}) \geq 0, \tag{2.4}$$

where $\Lambda = X^{\mathrm{T}} H(X) X$ and

$$\mathcal{T}(X) := \{S \mid X^{\mathrm{T}} S + S^{\mathrm{T}} X = 0\}. \tag{2.5}$$

(2) *Suppose that $X \in \mathbb{R}^{n\times p}$ satisfies* (2.1) *with a symmetric matrix Λ and* (2.4) *holds with a strict inequality for all $0 \neq S \in \mathcal{T}(X)$. Then X is a strict local minimizer for problem* (1.3).

Proof. It follows from Theorem 12.5 in [26] that the second-order necessary condition for X to be a local minimizer of (1.3) is

$$\left\langle S, \nabla^2_{XX} \mathcal{L}(X, \Lambda)[S] \right\rangle \geq 0 \quad \text{for all } S \in \mathcal{T}(X). \tag{2.6}$$

Using Lemma 2.1 and the fact that $\operatorname{tr}(X^{\mathrm{T}} \operatorname{Diag}(y) Z) = y^{\mathrm{T}} \operatorname{diag}(ZX^{\mathrm{T}})$ for all $X, Z \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^n$, we obtain

$$\begin{aligned}
\left\langle S, \nabla^2_{XX} \mathcal{L}(X, \Lambda)[S] \right\rangle &= \operatorname{tr}(S^{\mathrm{T}} \nabla^2 E(X)[S] - \Lambda S^{\mathrm{T}} S) \\
&= \operatorname{tr}\left(S^{\mathrm{T}} H(X) S + 2 S^{\mathrm{T}} \operatorname{Diag}\left(J \operatorname{diag}(SX^{\mathrm{T}})\right) X - \Lambda S^{\mathrm{T}} S\right) \\
&= \operatorname{tr}(S^{\mathrm{T}} H(X) S - \Lambda S^{\mathrm{T}} S) + 2 \operatorname{diag}(XS^{\mathrm{T}})^{\mathrm{T}} J \operatorname{diag}(XS^{\mathrm{T}}),
\end{aligned}$$

which together with (2.6) yields (2.4). The second part is a direct application of Theorem 12.6 in [26]. ∎

An equivalent formulation of the tangent space (2.5) is

$$\mathcal{T}(X) = \{S := XK + \mathbf{P}_X^{\perp} Z \mid K = -K^{\mathrm{T}} \in \mathbb{R}^{p\times p},\ Z \in \mathbb{R}^{n\times p}\}, \tag{2.7}$$

where $\mathbf{P}_X^{\perp} := I - XX^{\mathrm{T}}$. Hence, the second-order optimality conditions in Theorem 2.2 can be presented in terms of an arbitrary $Z \in \mathbb{R}^{n\times p}$ similar to the analysis of maximization of the sum of the trace ratio on the Stiefel manifold in [34, 35].

THEOREM 2.3.

(1) *Suppose that X is a local minimizer of problem* (1.3) *and $\epsilon_{xc}(\rho(X))$ is twice differentiable with respect to $\rho(X)$. Then for all $Z \in \mathbb{R}^{n\times p}$, it holds*

$$\begin{aligned}
\operatorname{tr}(Z^{\mathrm{T}} H(X) Z) + \operatorname{tr}(X^{\mathrm{T}} Z \Lambda Z^{\mathrm{T}} X) - \operatorname{tr}(Z^{\mathrm{T}} X \Lambda X^{\mathrm{T}} Z) - \operatorname{tr}(Z \Lambda Z^{\mathrm{T}}) \\
+ 2 \operatorname{diag}(XZ^{\mathrm{T}} \mathbf{P}_X^{\perp})^{\mathrm{T}} J \operatorname{diag}(XZ^{\mathrm{T}} \mathbf{P}_X^{\perp}) \geq 0.
\end{aligned} \tag{2.8}$$

(2) *Suppose that $X \in \mathbb{R}^{n\times p}$ satisfies* (2.1) *with a symmetric matrix Λ and* (2.8) *holds with a strict inequality for all $\mathbf{P}_X^{\perp} Z \neq 0$. Then X is a strict local minimizer for problem* (1.3).

Proof. Using (2.1) and the definition of $\mathbf{P}_X^\perp$, we obtain $\mathbf{P}_X^\perp \mathbf{P}_X^\perp = \mathbf{P}_X^\perp$, $\mathbf{P}_X^\perp X = 0$, and $\mathbf{P}_X^\perp H(X)X = 0$. For any $S = XK + \mathbf{P}_X^\perp Z$, it holds

$$\begin{aligned}\operatorname{tr}(S^{\mathrm{T}}H(X)S) &= \operatorname{tr}(K^{\mathrm{T}}X^{\mathrm{T}}H(X)XK) + \operatorname{tr}(Z^{\mathrm{T}}\mathbf{P}_X^\perp H(X)\mathbf{P}_X^\perp Z)\\ &= \operatorname{tr}(K^{\mathrm{T}}\Lambda K) + \operatorname{tr}(Z^{\mathrm{T}}H(X)Z) - \operatorname{tr}(Z^{\mathrm{T}}H(X)XX^{\mathrm{T}}Z)\\ &= \operatorname{tr}(K^{\mathrm{T}}\Lambda K) + \operatorname{tr}(Z^{\mathrm{T}}H(X)Z) - \operatorname{tr}(Z^{\mathrm{T}}X\Lambda X^{\mathrm{T}}Z). \qquad (2.9)\end{aligned}$$

It can be verified that $S^{\mathrm{T}}S = K^{\mathrm{T}}K + Z^{\mathrm{T}}\mathbf{P}_X^\perp Z$, which yields

$$\begin{aligned}\operatorname{tr}(\Lambda S^{\mathrm{T}}S) &= \operatorname{tr}(K^{\mathrm{T}}K\Lambda) + \operatorname{tr}(Z^{\mathrm{T}}Z\Lambda) - \operatorname{tr}(Z^{\mathrm{T}}XX^{\mathrm{T}}Z\Lambda)\\ &= \operatorname{tr}(K^{\mathrm{T}}\Lambda K) + \operatorname{tr}(Z\Lambda Z^{\mathrm{T}}) - \operatorname{tr}(X^{\mathrm{T}}Z\Lambda Z^{\mathrm{T}}X), \qquad (2.10)\end{aligned}$$

where the last equality holds because of $K = -K^{\mathrm{T}}$. Since it holds

$$\operatorname{diag}(XK^{\mathrm{T}}X^{\mathrm{T}}) = \frac{1}{2}(\operatorname{diag}(XK^{\mathrm{T}}X^{\mathrm{T}}) + \operatorname{diag}(XKX^{\mathrm{T}})) = \frac{1}{2}\operatorname{diag}(X(K+K^{\mathrm{T}})X^{\mathrm{T}}) = 0,$$

we obtain

$$\operatorname{diag}(XS^{\mathrm{T}}) = \operatorname{diag}(XK^{\mathrm{T}}X^{\mathrm{T}}) + \operatorname{diag}(XZ^{\mathrm{T}}\mathbf{P}_X^\perp) = \operatorname{diag}(XZ^{\mathrm{T}}\mathbf{P}_X^\perp),$$

which together with (2.9) and (2.10) gives (2.8). The proof of the second part follows directly from Theorem 2.2. ∎

2.3. Conditions for local minimizers to satisfy the KS equation. In this subsection, we establish a condition under which a local minimizer of (1.3) is a solution of a modification of the KS equation (1.5). Our discussion is restricted to a special exchange functional (the correlation term is ignored)

$$e^{\mathrm{T}}\epsilon_{xc}(\rho) = -\frac{3}{4}\gamma\rho^{\mathrm{T}}\rho^{\frac{1}{3}}, \qquad (2.11)$$

where $\gamma = 2\left(\frac{3}{\pi}\right)^{1/3}$ and $\rho^{\frac{1}{3}}$ denotes the componentwise cubic root of the vector ρ. The next result shows that the charge density ρ is bounded.

LEMMA 2.4. *Let $X \in \mathbb{R}^{n\times p}$ satisfy $X^{\mathrm{T}}X = I$ and ρ be defined by* (1.1). *We have*

$$0 \le \rho_i \le 1 \text{ for all } i = 1, \dots, n. \qquad (2.12)$$

Proof. The inequality (2.12) holds from $X^{\mathrm{T}}X = I$ and the fact that $\rho_i = \sum_{j=1}^p X_{ij}^2$ for all $i = 1, \dots, n$. ∎

Our analysis relies on the gap between the pth and $(p+1)$st smallest eigenvalues of $H(X)$.

Assumption 2.5. Let $\lambda_1 \le \cdots \le \lambda_p \le \lambda_{p+1} \le \cdots \le \lambda_n$ be the eigenvalues of a given symmetric matrix $H \in \mathbb{R}^{n\times n}$. There exists a positive constant δ such that $\lambda_{p+1} - \lambda_p \ge \delta$.

Note that $E(X)$ may not be second-order differentiable since some components $\rho_i(X)$ can be zero. Let $\mathcal{I}$ be the collection of indices of the nonzero components of $\rho(X)$, i.e.,

$$\mathcal{I} = \{i \mid \rho_i(X) \ne 0, i = 1, \dots, n\}. \qquad (2.13)$$

Then the complement set $\bar{\mathcal{I}}$ of $\mathcal{I}$ is the set of indices of the zero components of $\rho(X)$. Let r be the cardinality of $\mathcal{I}$. We have $r \ge p$ by the orthogonality of X. If

$\mathcal{I} = \{\alpha_1, \ldots, \alpha_r\}$, we define the submatrices $X_{\mathcal{I}}$ and $L_{\mathcal{I}\mathcal{I}}$ as

$$X_{\mathcal{I}} = \begin{pmatrix} X_{\alpha_1,1}, \ldots, X_{\alpha_1,p} \\ \cdots \\ X_{\alpha_r,1}, \ldots, X_{\alpha_r,p} \end{pmatrix} \text{ and } L_{\mathcal{I}\mathcal{I}} = \begin{pmatrix} L_{\alpha_1,\alpha_1}, \ldots, L_{\alpha_1,\alpha_r} \\ \cdots \\ L_{\alpha_r,\alpha_1}, \ldots, L_{\alpha_r,\alpha_r} \end{pmatrix}.$$

The notation $(V_{ion})_{\mathcal{I}\mathcal{I}}$, $L^{\dagger}_{\mathcal{I}\mathcal{I}}$, $H_{\mathcal{I}\mathcal{I}}(X)$, and $\Lambda_{\mathcal{I}\mathcal{I}}$ are defined similarly to $L_{\mathcal{I}\mathcal{I}}$.

The following theorem shows that a local minimizer X^* of the KS total energy minimization (1.3) is a solution of KS equation (1.5) if all rows of X^* are nonzero and Assumption 2.5 holds with a sufficiently large gap δ.

THEOREM 2.6. *Suppose that X^* is a local minimizer of* (1.3) *using* (2.11) *and $\Lambda^* = (X^*)^{\top} H(X^*) X^*$ is a diagonal matrix. Let $\mathcal{I}^*$ be the index set of X^* defined as* (2.13). *If Assumption* 2.5 *holds at $H(X^*)$ with a constant δ satisfying*

$$\delta > 2\left(||L^{\dagger}||_2 - \frac{\gamma}{3}\right), \tag{2.14}$$

then it holds

$$\begin{aligned} H_{\mathcal{I}^*\mathcal{I}^*}(X^*) X^*_{\mathcal{I}^*} &= X^*_{\mathcal{I}^*} \Lambda^*, \\ (X^*_{\mathcal{I}^*})^{\mathrm{T}} X^*_{\mathcal{I}^*} &= I, \end{aligned} \tag{2.15}$$

and the diagonal of Λ^* consists of the p smallest eigenvalues of $H_{\mathcal{I}^*\mathcal{I}^*}(X^*)$.

Proof. It can be verified that X^* is a local minimizer of the restricted problem

$$\begin{aligned} \min_{X \in \mathbb{R}^{n\times p}} \quad & E(X) \\ \text{s.t.} \quad & X^{\mathrm{T}} X = I,\ X_{\bar{\mathcal{I}}^*} = 0. \end{aligned} \tag{2.16}$$

Hence, $X^*_{\mathcal{I}^*}$ is a local minimizer of the reduced problem

$$\begin{aligned} \min_{\hat{X} \in \mathbb{R}^{r\times p}} \quad & \hat{E}(\hat{X}) := \frac{1}{4}\mathrm{tr}(\hat{X}^{\mathrm{T}} L_{\mathcal{I}^*\mathcal{I}^*} \hat{X}) + \frac{1}{2}\mathrm{tr}(\hat{X}^{\mathrm{T}} (V_{ion})_{\mathcal{I}^*\mathcal{I}^*} \hat{X}) \\ & \qquad + \frac{1}{4}\rho(\hat{X})^{\top} L^{\dagger}_{\mathcal{I}^*\mathcal{I}^*} \rho(\hat{X}) - \frac{3}{4}\gamma \rho(\hat{X})^{\mathrm{T}} \rho(\hat{X})^{\frac{1}{3}} \\ \text{s.t.} \quad & \hat{X}^{\mathrm{T}} \hat{X} = I, \end{aligned} \tag{2.17}$$

The structure of the energy functional $E(X)$ implies $\nabla \hat{E}(X^*_{\mathcal{I}^*}) = H_{\mathcal{I}^*\mathcal{I}^*}(X^*) X^*_{\mathcal{I}^*}$ and $(X^*_{\mathcal{I}^*})^{\mathrm{T}} H_{\mathcal{I}^*\mathcal{I}^*}(X^*_{\mathcal{I}^*}) X^*_{\mathcal{I}^*} = \Lambda^*$. These facts together with the first-order optimality of (2.17) at $X^*_{\mathcal{I}^*}$ yield (2.15).

It is obvious that the diagonal entries of Λ^* are the eigenvalues of $H_{\mathcal{I}^*\mathcal{I}^*}(X^*)$. Suppose that they are not the p smallest eigenvalues of $H_{\mathcal{I}^*\mathcal{I}^*}(X^*)$. For convenience, we denote the eigenvalues of $H_{\mathcal{I}^*\mathcal{I}^*}(X^*)$ in an ascending order as $\hat{\lambda}_1 \le \cdots \le \hat{\lambda}_r$ and their corresponding eigenvectors are u_i, $i = 1, \ldots, r$, where $r = |\mathcal{I}^*|$. Let x_i, $1 \le i \le p$, be the ith column for $X^*_{\mathcal{I}^*}$. Without loss of generality, let x_1 be associated with an eigenvalue greater than $\hat{\lambda}_p$, and u_i $(i \le p)$ be an eigenvector associated with an eigenvalue less than or equal to $\hat{\lambda}_p$ but not be a column of $X^*_{\mathcal{I}^*}$. Assumption 2.5 implies that $u_i \notin \mathrm{span}\{X^*_{\mathcal{I}^*}\}$. Let V be a matrix whose columns satisfy

$$v_j = \begin{cases} u_i & \text{if } j = 1, \\ x_j & \text{if } j = 2, \ldots, p. \end{cases}$$

The function $\hat{E}(\hat{X})$ is twice differentiable at $X^*_{\mathcal{I}^*}$ according to the definition of $\mathcal{I}^*$.

Therefore, an application of Theorem 2.3 gives

$$\begin{aligned}\Delta :=\ & \mathrm{tr}(V^{\mathrm{T}}H_{\mathcal{I}^*\mathcal{I}^*}(X^*_{\mathcal{I}^*})V) + \mathrm{tr}((X^*_{\mathcal{I}^*})^{\mathrm{T}}V\Lambda^*V^{\mathrm{T}}X^*_{\mathcal{I}^*}) \\ & - \mathrm{tr}(V^{\mathrm{T}}X^*_{\mathcal{I}^*}\Lambda^*(X^*_{\mathcal{I}^*})^{\mathrm{T}}V) - \mathrm{tr}(V\Lambda^*V^{\mathrm{T}}) \\ & + 2\,\mathrm{diag}(X^*_{\mathcal{I}^*}V^{\mathrm{T}}\mathbf{P}^{\perp}_{X^*_{\mathcal{I}^*}})^{\mathrm{T}}\left(L^{\dagger}_{\mathcal{I}^*\mathcal{I}^*} - \frac{\gamma}{3}\mathrm{Diag}\left(\rho(X^*_{\mathcal{I}^*})^{-\frac{2}{3}}\right)\right)\mathrm{diag}(X^*_{\mathcal{I}^*}V^{\mathrm{T}}\mathbf{P}^{\perp}_{X^*_{\mathcal{I}^*}}) \\ & \geq 0.\end{aligned} \tag{2.18}$$

It follows from V is an orthonormal eigenbasis of $H_{\mathcal{I}^*\mathcal{I}^*}(X^*_{\mathcal{I}^*})$ and Assumption 2.5 that

$$\mathrm{tr}(V^{\mathrm{T}}H_{\mathcal{I}^*\mathcal{I}^*}(X^*)V) - \mathrm{tr}((X^*_{\mathcal{I}^*})^{\mathrm{T}}H_{\mathcal{I}^*\mathcal{I}^*}(X^*)X^*_{\mathcal{I}^*}) \leq \hat{\lambda}_i - \hat{\lambda}_{p+1} \leq -\delta. \tag{2.19}$$

Since $u_i \notin \mathrm{span}\{X^*_{\mathcal{I}^*}\}$, we obtain

$$(X^*_{\mathcal{I}^*})^{\mathrm{T}}V = V^{\mathrm{T}}X^*_{\mathcal{I}^*} = I - e_1e_1^{\mathrm{T}}, \tag{2.20}$$

$$X^*_{\mathcal{I}^*}V^{\mathrm{T}}\mathbf{P}^{\perp}_{X^*_{\mathcal{I}^*}} = x_1u_i^{\mathrm{T}}, \tag{2.21}$$

which further give

$$\begin{aligned}\Delta &= \mathrm{tr}(V^{\mathrm{T}}H_{\mathcal{I}^*\mathcal{I}^*}(X^*_{\mathcal{I}^*})V) - \mathrm{tr}(\Lambda^*) \\ &\quad + 2\,\mathrm{diag}(x_1u_i^{\mathrm{T}})^{\mathrm{T}}\left(L^{\dagger}_{\mathcal{I}^*\mathcal{I}^*} - \frac{\gamma}{3}\mathrm{Diag}\left(\rho(X^*_{\mathcal{I}^*})^{-\frac{2}{3}}\right)\right)\mathrm{diag}(x_1u_i^{\mathrm{T}}) \\ &\leq -\delta + 2\max\left\{\lambda_{\max}\left(L^{\dagger}_{\mathcal{I}^*\mathcal{I}^*} - \frac{\gamma}{3}\mathrm{Diag}\left(\rho(X^*_{\mathcal{I}^*})^{-\frac{2}{3}}\right)\right), 0\right\} \\ &\leq -\delta + 2\max\left\{\lambda_{\max}\left(L^{\dagger}_{\mathcal{I}^*\mathcal{I}^*} - \frac{\gamma}{3}I\right), 0\right\} \\ &\leq -\delta + 2\max\left\{\left(||L^{\dagger}_{\mathcal{I}^*\mathcal{I}^*}||_2 - \frac{\gamma}{3}\right), 0\right\} \\ &< 0,\end{aligned} \tag{2.22}$$

where the first inequality uses (2.19) and the fact that $||\,\mathrm{diag}(x_1u_i^{\mathrm{T}})||_2^2 \leq 1$, the second inequality follows from $\rho \in [0,1]$, the third inequality uses the fact that $||L^{\dagger}_{\mathcal{I}^*\mathcal{I}^*}||_2 \leq ||L^{\dagger}||_2$ since the largest/smallest eigenvalue of a matrix is no less/greater than the largest/smallest eigenvalue of its principal submatrix, and the last inequality (2.22) is due to (2.14). However, (2.22) is a contradiction to (2.18). This completes the proof. ∎

2.4. Conditions for global minimizers to satisfy the KS equation. In this subsection, we consider whether a global minimizer of (1.3) is a solution of the KS equation (1.5) under the exchange functional (2.11). We first show the following inequality.

LEMMA 2.7. *It holds for all* $a, b \in [0,1]$ *that* $(a-b)^2(3a^2+2ab+b^2) = 3a^4 - 4a^3b + b^4 \geq \frac{2}{3}(a^3-b^3)^2$.

Proof. The inequality holds for $a = 0$ or $b = 0$. Consider the case $a \geq b > 0$. Introducing the variable $t = b/a \in (0,1]$ yields $a^4(3-4t+t^4) - \frac{2}{3}a^6(1-t^3)^2 \geq a^6f(t)$, where $f(t) = 3 - 4t + t^4 - \frac{2}{3}(1-t^3)^2$. Since $f'(t) = (t^3-1)(4-4t^2) \leq 0$ for all $t \in [0,1]$, we have $f(t) \geq f(1) = 0$ for all $t \in [0,1]$, and then the inequality is proved. The case $b \geq a > 0$ can be proved in a similar fashion. ∎

The next theorem establishes the equivalence based on estimating the difference of total energy function values.

THEOREM 2.8. *Suppose that X^* is a global minimizer of* (1.3) *using* (2.11). *If Assumption* 2.5 *holds at $H(X^*)$ with a constant δ satisfying*

$$\delta > p\left(||L^\dagger||_2 - \frac{\gamma}{3}\right), \tag{2.23}$$

then X^* must be an orthonormal eigenbasis of $H(X^*)$ corresponding to its p smallest eigenvalues, namely, a solution of the KS equation (1.5).

Proof. Suppose that X^* is not but Y is an orthonormal eigenbasis of $H(X^*)$ corresponding to its p smallest eigenvalues. Since X^* must be an orthonormal eigenbasis of $H(X^*)$ and using Assumption 2.5, we have

$$\begin{aligned}\Delta H(Y, X^*) &:= \mathrm{tr}(Y^{\mathrm{T}} H(X^*) Y) - \mathrm{tr}((X^*)^{\mathrm{T}} H(X^*) X^*) \\ &\le \lambda_p(H(X^*)) - \lambda_{p+1}(H(X^*)) \le -\delta.\end{aligned} \tag{2.24}$$

Applying Lemmas 2.4 and 2.7 gives

$$\sum_{i=1}^{n}\left(\rho(Y)_i^{\frac{1}{3}} - \rho(X^*)_i^{\frac{1}{3}}\right)^2\left(3\rho(Y)_i^{\frac{2}{3}} + 2\rho(Y)_i^{\frac{1}{3}}\rho(X^*)_i^{\frac{1}{3}} + \rho(X^*)_i^{\frac{2}{3}}\right) \ge \frac{2}{3}||\rho(Y) - \rho(X^*)||_2^2. \tag{2.25}$$

It follows from Lemma 2.4 that

$$\begin{aligned}\|\rho(Y) - \rho(X^*)\|^2 &\le (1 - \rho(Y))^{\mathrm{T}}\rho(X^*) + (1 - \rho(X^*))^{\mathrm{T}}\rho(Y) \\ &\le 1^{\mathrm{T}}\rho(X^*) + 1^{\mathrm{T}}\rho(Y) = \mathrm{tr}(XX^{\mathrm{T}}) + \mathrm{tr}(YY^{\mathrm{T}}) \\ &= 2p.\end{aligned} \tag{2.26}$$

Using the relationship $\mathrm{tr}(Y^{\mathrm{T}}\mathrm{Diag}(L^\dagger\rho(X^*))Y) = \rho(Y)^{\mathrm{T}}L^\dagger\rho(X^*)$, the inequalities (2.24), (2.25), and (2.26), and the assumption (2.23), we obtain

$$\begin{aligned}\Delta E(Y, X^*) &= E(Y) - E(X^*) \\ &= \frac{1}{2}\Delta H(Y, X^*) + \frac{1}{4}\left(\rho(Y)^{\mathrm{T}}L^\dagger\rho(Y) - \rho(X^*)^{\mathrm{T}}L^\dagger\rho(X^*)\right) \\ &\quad - \frac{3\gamma}{8}\left(\rho(Y)^{\mathrm{T}}\rho(Y)^{\frac{1}{3}} - \rho(X^*)^{\mathrm{T}}\rho(X^*)^{\frac{1}{3}}\right) \\ &\quad - \frac{1}{2}\mathrm{tr}(Y^{\mathrm{T}}\mathrm{Diag}(L^\dagger\rho(X^*) - \gamma\rho(X^*)^{\frac{1}{3}})Y) \\ &\quad + \frac{1}{2}\mathrm{tr}(X^{\mathrm{T}}\mathrm{Diag}(L^\dagger\rho(X^*) - \gamma\rho(X^*)^{\frac{1}{3}})X^*) \\ &= \frac{1}{2}\Delta H(Y, X^*) + \frac{1}{4}\left(\rho(Y)^{\mathrm{T}}L^\dagger\rho(Y) - \rho(X^*)^{\mathrm{T}}L^\dagger\rho(X^*)\right) \\ &\quad - \frac{3\gamma}{8}\left(\rho(Y)^{\mathrm{T}}\rho(Y)^{\frac{1}{3}} - \rho(X^*)^{\mathrm{T}}\rho(X^*)^{\frac{1}{3}}\right) \\ &\quad - \frac{1}{2}\left(\rho(Y)^{\mathrm{T}}L^\dagger\rho(X^*) - \rho(X^*)^{\mathrm{T}}L^\dagger\rho(X^*)\right) \\ &\quad + \frac{1}{2}\gamma\left(\rho(Y)^{\mathrm{T}}\rho(X^*)^{\frac{1}{3}} - \rho(X^*)^{\mathrm{T}}\rho(X^*)^{\frac{1}{3}}\right) \\ &= \frac{1}{2}\Delta H(Y, X^*) + \frac{1}{4}(\rho(Y) - \rho(X^*))^{\mathrm{T}}L^\dagger(\rho(Y) - \rho(X^*)) \\ &\quad - \frac{\gamma}{8}\sum_{i=1}^{n}\left(\rho(Y)_i^{\frac{1}{3}} - \rho(X^*)_i^{\frac{1}{3}}\right)^2\left(3\rho(Y)_i^{\frac{2}{3}} + 2\rho(Y)_i^{\frac{1}{3}}\rho(X^*)_i^{\frac{1}{3}} + \rho(X^*)_i^{\frac{2}{3}}\right)\end{aligned}$$

$$\begin{aligned}
&\leq -\frac{\delta}{2} + \left(\frac{||L^\dagger||_2}{4} - \frac{\gamma}{12}\right) ||\rho(Y) - \rho(X^*)||_2^2 \\
&\leq -\frac{\delta}{2} + \left(\frac{||L^\dagger||_2}{4} - \frac{\gamma}{12}\right) (2p) \\
&< 0,
\end{aligned}$$

which is a contradiction to the fact that X^* is a global minimizer. This completes the proof. ∎

Remark 2.9. When the exchange correlation function $\epsilon_{xc}(\rho)$ is equal to zero, our condition (2.23) becomes $\delta > p||L^\dagger||_2$, which is much weaker than the condition $\delta > 12p\sqrt{n}||L^\dagger||_2$ in Theorem 1 of [25].

2.5. Lower bounds for the charge density of local minimizers. The exchange correlation energy functional is twice differentiable if all components of $\rho(X)$ are positive. However, the second-order derivative may not be bounded at an arbitrary point X. In this subsection, we provide a few lower bounds for the charge density at certain types of local minimizers. These properties are useful for our analysis on the KS equation.

Traditionally, a point x^* is called a strong local minimizer [2, 16] of a function $f : \mathbb{R}^n \mapsto \mathbb{R}$, if there exist a constant $\kappa > 0$ and a neighborhood U of x^* such that the inequality

$$f(x) \geq f(x^*) + \kappa||x - x^*||_2^2 \tag{2.27}$$

holds for any $x \in U$. Here, we define a strong local minimizer based on the second-order optimality conditions.

DEFINITION 2.10. *A point X^* is called a strong local minimizer of* (1.3) *using* (2.11) *if and only if $X^*_{\mathcal{I}^*}$ is a local minimizer of* (2.17) *and there exists a constant $\kappa > 0$ such that for all $Z \in \mathbb{R}^{n\times p}$,*

$$\begin{aligned}
&\operatorname{tr}(Z^{\mathrm{T}} H_{\mathcal{I}^*\mathcal{I}^*}(X^*_{\mathcal{I}^*})Z) + \operatorname{tr}((X^*_{\mathcal{I}^*})^{\mathrm{T}} Z\Lambda^* Z^{\mathrm{T}} X^*_{\mathcal{I}^*}) - \operatorname{tr}(Z^{\mathrm{T}} X^*_{\mathcal{I}^*}\Lambda^*(X^*_{\mathcal{I}^*})^{\mathrm{T}} Z) \\
&\quad - \operatorname{tr}(Z\Lambda^* Z^{\mathrm{T}}) + 2\operatorname{diag}((X^*_{\mathcal{I}^*})Z^{\mathrm{T}}\mathbf{P}^{\perp}_{X^*_{\mathcal{I}^*}})^{\mathrm{T}} \left(L^\dagger_{\mathcal{I}^*\mathcal{I}^*} - \frac{\gamma}{3}\operatorname{Diag}\left(\rho(X^*_{\mathcal{I}^*})^{-\frac{2}{3}}\right)\right) \\
&\quad \cdot \operatorname{diag}((X^*_{\mathcal{I}^*})Z^{\mathrm{T}}\mathbf{P}^{\perp}_{X^*_{\mathcal{I}^*}}) \geq \kappa\|Z\|_F^2,
\end{aligned} \tag{2.28}$$

where $\Lambda^* = (X^*_{\mathcal{I}^*})^{\mathrm{T}} H_{\mathcal{I}^*\mathcal{I}^*}(X^*)X^*_{\mathcal{I}^*}$ and $\mathcal{I}^*$ is the index set of X^* defined as (2.13).

Our condition (2.28) is weaker than (2.27) applied to problem (1.3) when the total energy $E(X)$ is twice differentiable. The next result shows that the charge densities at a strong local minimizer are bounded below uniformly if they are positive.

THEOREM 2.11. *Suppose that L is positive semidefinite and X^* is a strong local minimizer of* (1.3) *satisfying Definition* 2.10. *Let*

$$\bar{c} := \min\{1, c_1, \ldots, c_n\} \text{ and } c_i := \min_{j\neq i}\left(\frac{\gamma}{3(L^\dagger_{ii} - 2L^\dagger_{ij} + L^\dagger_{jj})}\right)^{\frac{3}{2}}. \tag{2.29}$$

Then it holds

$$\text{for any } i \in \{1, 2, \ldots, n\}, \qquad \rho_i(X^*) \in [0, \bar{c}) \quad \Rightarrow \quad \rho_i(X^*) = 0. \tag{2.30}$$

Proof. For convenience, we denote $\rho^*_{\mathcal{I}^*} = \rho(X^*_{\mathcal{I}^*})$. If there exists a row j in $X^*_{\mathcal{I}^*}$ such that either 1 or -1 is an entry of this row, then this row has only one nonzero entry according to the orthonormality of $X^*_{\mathcal{I}^*}$. Hence, $(\rho^*_{\mathcal{I}^*})_j = 1$ and (2.30) holds at j.

We next consider the components in the set $\mathcal{J} := \{j \mid j \in \mathcal{I}^* \text{ and } |(X^*_{\mathcal{I}^*})_{js}| < 1,\ s = 1, \ldots, p\}$.

For any given $j \in \mathcal{J}$, there exists a nonzero entry, denoted as $(X^*_{\mathcal{I}^*})_{js}$, in the jth row of $X^*_{\mathcal{I}^*}$. Since $|(X^*_{\mathcal{I}^*})_{js}| < 1$, there exists at least another nonzero entry, denoted as $(X^*_{\mathcal{I}^*})_{is}$, in the sth column of $X^*_{\mathcal{I}^*}$ due to the orthonormality of $X^*_{\mathcal{I}^*}$. For simplicity, let x_l, $l = 1, \ldots, p$, be the lth column of $X^*_{\mathcal{I}^*}$ and set $r = |\mathcal{I}^*|$, $x_{js} = (X^*_{\mathcal{I}^*})_{js}$ and $x_{is} = (X^*_{\mathcal{I}^*})_{is}$. Define a vector $z \in \mathbb{R}^r$ whose lth component ($l = 1, \ldots, p$) is

$$z_l = \begin{cases} \frac{x_{is}}{\sqrt{x_{is}^2 + x_{js}^2}} & \text{if } l = j, \\ \frac{-x_{js}}{\sqrt{x_{is}^2 + x_{js}^2}} & \text{if } l = i, \\ 0 & \text{otherwise.} \end{cases} \tag{2.31}$$

A short calculation gives $||z||_2 = 1$, $z^{\mathrm{T}} x_s = 0$, and

$$\operatorname{diag}(z x_s^{\mathrm{T}}) = \frac{x_{is} x_{js}}{\sqrt{x_{is}^2 + x_{js}^2}} e_{(j,-i)}, \tag{2.32}$$

where $e_{(j,-i)} \in \mathbb{R}^r$ has 1 on its jth entry, -1 on its ith entry, and 0 elsewhere.

For $a \in [0, 1]$, let $Z_a \in \mathbb{R}^{n \times p}$ be a matrix whose sth column is $az + \sqrt{1 - a^2} x_s$ and all other columns are zero. Without loss of generality, let $\hat{\lambda}_1 \le \cdots \le \hat{\lambda}_r$ be the eigenvalues of $H_{\mathcal{I}^*\mathcal{I}^*}(X^*)$ in ascending order, and x_s be an eigenvector of $H_{\mathcal{I}^*\mathcal{I}^*}(X^*)$ associated with $\hat{\lambda}_s$, $s \in \{1, \ldots, r\}$. Then, we obtain

$$\operatorname{tr}(Z_a^{\mathrm{T}} H_{\mathcal{I}^*\mathcal{I}^*}(X^*) Z_a) \le a^2 \hat{\lambda}_r + (1 - a^2) \hat{\lambda}_s, \tag{2.33}$$

$$\operatorname{tr}(Z_a \Lambda^* Z_a^{\mathrm{T}}) = \operatorname{tr}(\Lambda^* Z_a^{\mathrm{T}} Z_a) = \hat{\lambda}_s, \tag{2.34}$$

which yield

$$\begin{aligned} &\operatorname{tr}(Z_a^{\mathrm{T}} H_{\mathcal{I}^*\mathcal{I}^*}(X^*) Z_a) - \operatorname{tr}(Z_a \Lambda^* Z_a^{\mathrm{T}}) \\ &\quad \le a^2 \hat{\lambda}_r + (1 - a^2) \hat{\lambda}_s - \hat{\lambda}_s = a^2 (\hat{\lambda}_r - \hat{\lambda}_s). \end{aligned} \tag{2.35}$$

The definition of Z_a gives

$$(Z_a^{\mathrm{T}} X^*_{\mathcal{I}^*})_{pq} = \begin{cases} a z^{\mathrm{T}} x_q & \text{if } p = s,\ q \ne s, \\ \sqrt{1 - a^2} & \text{if } p = s,\ q = s, \\ 0 & \text{otherwise.} \end{cases} \tag{2.36}$$

Hence, we have

$$\begin{aligned} \operatorname{tr}((X^*_{\mathcal{I}^*})^{\mathrm{T}} Z_a \Lambda^* Z_a^{\mathrm{T}} X^*_{\mathcal{I}^*}) &= \operatorname{tr}(\Lambda^* Z_a^{\mathrm{T}} X^*_{\mathcal{I}^*} (X^*_{\mathcal{I}^*})^{\mathrm{T}} Z_a) \\ &= \hat{\lambda}_s \left(\sum_{q=1, q \ne s}^{p} a^2 (z^{\mathrm{T}} x_q)^2 + (1 - a^2) \right) \\ &= \hat{\lambda}_s \left(\sum_{q=1}^{p} a^2 (z^{\mathrm{T}} x_q)^2 + (1 - a^2) - a^2 (z^{\mathrm{T}} x_s)^2 \right) \\ &= \hat{\lambda}_s (1 + a^2 ||z^{\mathrm{T}} X^*_{\mathcal{I}^*}||_2^2 - a^2) = a^2 \hat{\lambda}_s ||z^{\mathrm{T}} X^*_{\mathcal{I}^*}||_2^2 + (1 - a^2) \hat{\lambda}_s \end{aligned} \tag{2.37}$$

and

$$\begin{aligned}\operatorname{tr}(Z_a^{\mathrm{T}} X_{\mathcal{I}^*}^* \Lambda^* (X_{\mathcal{I}^*}^*)^{\mathrm{T}} Z_a) &= \left(\sum_{q=1, q\neq s}^{p} a^2 (z^{\mathrm{T}} x_q)^2 \hat{\lambda}_q + (1-a^2)\hat{\lambda}_s\right) \\ &\geq \left(\sum_{q=1, q\neq s}^{p} a^2 (z^{\mathrm{T}} x_q)^2 \hat{\lambda}_1 + (1-a^2)\hat{\lambda}_s\right) \\ &= \left(\sum_{q=1}^{p} a^2 (z^{\mathrm{T}} x_q)^2 \hat{\lambda}_1 + (1-a^2)\hat{\lambda}_s\right) \\ &= a^2 \hat{\lambda}_1 ||z^{\mathrm{T}} X_{\mathcal{I}^*}^*||_2^2 + (1-a^2)\hat{\lambda}_s. \end{aligned} \tag{2.38}$$

Combining (2.37) and (2.38) yields

$$\begin{aligned}&\operatorname{tr}((X_{\mathcal{I}^*}^*)^{\mathrm{T}} Z_a \Lambda^* Z_a^{\mathrm{T}} X_{\mathcal{I}^*}^*) - \operatorname{tr}(Z_a^{\mathrm{T}} X_{\mathcal{I}^*}^* \Lambda^* (X_{\mathcal{I}^*}^*)^{\mathrm{T}} Z_a) \\ &\quad \leq (a^2 \hat{\lambda}_s ||z^{\mathrm{T}} X_{\mathcal{I}^*}^*||_2^2 + (1-a^2)\hat{\lambda}_s) - (a^2 \hat{\lambda}_1 ||z^{\mathrm{T}} X_{\mathcal{I}^*}^*||_2^2 + (1-a^2)\hat{\lambda}_s) \\ &\quad = a^2(\hat{\lambda}_s - \hat{\lambda}_1)||z^{\mathrm{T}} X_{\mathcal{I}^*}^*||_2^2 \\ &\quad \leq a^2(\hat{\lambda}_s - \hat{\lambda}_1). \end{aligned} \tag{2.39}$$

The equality (2.32) gives

$$\operatorname{diag}(X_{\mathcal{I}^*}^* Z_a^{\mathrm{T}} \mathbf{P}_{X_{\mathcal{I}^*}^*}^{\perp}) = a \operatorname{diag}(x_s z^{\mathrm{T}}) = \frac{a x_{is} x_{js}}{\sqrt{x_{is}^2 + x_{js}^2}} e_{(j,-i)}. \tag{2.40}$$

Let Z_a with $a = \sqrt{\frac{\kappa}{\hat{\lambda}_r - \hat{\lambda}_1}}$. Using (2.35) and (2.39), we have

$$\begin{aligned}&\operatorname{tr}(Z_a^{\mathrm{T}} H_{\mathcal{I}^*\mathcal{I}^*}(X_{\mathcal{I}^*}^*) Z_a) + \operatorname{tr}((X_{\mathcal{I}^*}^*)^{\mathrm{T}} Z_a \Lambda_{\mathcal{I}^*} Z_a^{\mathrm{T}} X_{\mathcal{I}^*}^*) \\ &\quad - \operatorname{tr}(Z_a^{\mathrm{T}} X_{\mathcal{I}^*}^* \Lambda_{\mathcal{I}^*} (X_{\mathcal{I}^*}^*)^{\mathrm{T}} Z_a) - \operatorname{tr}(Z_a \Lambda_{\mathcal{I}^*} Z_a^{\mathrm{T}}) \leq a^2(\hat{\lambda}_r - \hat{\lambda}_1) = \kappa. \end{aligned} \tag{2.41}$$

It follows from our definition of strong local minimizers that

$$\begin{aligned}&\operatorname{tr}(Z_a^{\mathrm{T}} H_{\mathcal{I}^*\mathcal{I}^*}(X^*) Z_a) + \operatorname{tr}((X_{\mathcal{I}^*}^*)^{\mathrm{T}} Z_a \Lambda^* Z_a^{\mathrm{T}} X_{\mathcal{I}^*}^*) \\ &\quad - \operatorname{tr}(Z_a^{\mathrm{T}} X_{\mathcal{I}^*}^* \Lambda^* (X_{\mathcal{I}^*}^*)^{\mathrm{T}} Z_a) - \operatorname{tr}(Z_a \Lambda^* Z_a^{\mathrm{T}}) \\ &\quad + 2 \operatorname{diag}(X_{\mathcal{I}^*}^* Z_a^{\mathrm{T}} \mathbf{P}_X^{\perp})^{\mathrm{T}} \left(L_{\mathcal{I}^*\mathcal{I}^*}^{\dagger} - \frac{\gamma}{3} \operatorname{Diag}\left((\rho_{\mathcal{I}^*}^*)^{-\frac{2}{3}}\right)\right) \operatorname{diag}(X_{\mathcal{I}^*}^* Z_a^{\mathrm{T}} \mathbf{P}_{X_{\mathcal{I}^*}^*}^{\perp}) \geq \kappa, \end{aligned} \tag{2.42}$$

which together with (2.41) gives

$$\operatorname{diag}(X_{\mathcal{I}^*}^* Z_a^{\mathrm{T}} \mathbf{P}_{X_{\mathcal{I}^*}^*}^{\perp})^{\mathrm{T}} \left(L_{\mathcal{I}^*\mathcal{I}^*}^{\dagger} - \frac{\gamma}{3} \operatorname{Diag}\left((\rho_{\mathcal{I}^*}^*)^{-\frac{2}{3}}\right)\right) \operatorname{diag}(X_{\mathcal{I}^*}^* Z_a^{\mathrm{T}} \mathbf{P}_{X_{\mathcal{I}^*}^*}^{\perp}) \geq 0. \tag{2.43}$$

Substituting (2.40) into (2.43), we obtain

$$e_{(j,-i)}^{\mathrm{T}} \left(L_{\mathcal{I}^*\mathcal{I}^*}^{\dagger} - \frac{\gamma}{3} \operatorname{Diag}\left((\rho_{\mathcal{I}^*}^*)^{-\frac{2}{3}}\right)\right) e_{(j,-i)} \geq 0. \tag{2.44}$$

Expending the terms of (2.44) yields

$$(L_{\mathcal{I}^*\mathcal{I}^*}^{\dagger})_{jj} - 2(L_{\mathcal{I}^*\mathcal{I}^*}^{\dagger})_{ji} + (L_{\mathcal{I}^*\mathcal{I}^*}^{\dagger})_{ii} - \frac{\gamma}{3}(\rho_{\mathcal{I}^*}^*)_j^{-\frac{2}{3}} - \frac{\gamma}{3}(\rho_{\mathcal{I}^*}^*)_i^{-\frac{2}{3}} \geq 0, \tag{2.45}$$

which implies

$$(L^\dagger_{\mathcal{I}^*\mathcal{I}^*})_{jj} - 2(L^\dagger_{\mathcal{I}^*\mathcal{I}^*})_{ji} + (L^\dagger_{\mathcal{I}^*\mathcal{I}^*})_{ii} \geq \frac{\gamma}{3}(\rho^*_{\mathcal{I}^*})_j^{-\frac{2}{3}}. \tag{2.46}$$

Therefore, we obtain

$$(\rho^*_{\mathcal{I}^*})_j \geq \left(\frac{\gamma}{3((L^\dagger_{\mathcal{I}^*\mathcal{I}^*})_{jj} - 2(L^\dagger_{\mathcal{I}^*\mathcal{I}^*})_{ji} + (L^\dagger_{\mathcal{I}^*\mathcal{I}^*})_{ii})}\right)^{\frac{3}{2}} \geq c_j, \tag{2.47}$$

where c_j is defined in (2.29). Similarly, we can prove (2.47) holds for any $j \in \mathcal{J}$. This completes the proof. ∎

3. Analysis of the KS equation.

3.1. Formulating the KS equation as a fixed point map. The KS equation (1.5) constitutes a nonlinear system with respect to X. Note that the Hamiltonian matrix (1.4) is a symmetric matrix function with respect to ρ as

$$\hat{H}(\rho) := \frac{1}{2}L + V_{ion} + \mathrm{Diag}(L^\dagger\rho) + \mathrm{Diag}(\mu_{xc}(\rho)^{\mathrm{T}}e), \tag{3.1}$$

and the KS equation becomes

$$\begin{cases} \hat{H}(\rho)X = X\Lambda, \\ X^{\mathrm{T}}X = I, \end{cases} \tag{3.2}$$

where $X \in \mathbb{R}^{n\times p}$ and $\Lambda \in \mathbb{R}^{p\times p}$ is a diagonal matrix consisting of the p smallest eigenvalues of $\hat{H}(\rho)$. The eigenvalue decomposition of $\hat{H}(\rho)$ is determined once ρ is given. Hence, we can write X as $X(\rho)$ to reflect the dependence on ρ and the KS equation (1.5) can be viewed as a system of nonlinear equations with respect to the charge density ρ as

$$\rho = \mathrm{diag}(X(\rho)X(\rho)^{\mathrm{T}}). \tag{3.3}$$

Alternatively, the function

$$V := \mathcal{V}(\rho) = L^\dagger\rho + \mu_{xc}(\rho)^{\mathrm{T}}e \tag{3.4}$$

is called the potential and the Hamiltonian matrix $\hat{H}(\rho)$, by slight abuse of notation, can be expressed as

$$H(V) := \frac{1}{2}L + V_{ion} + \mathrm{Diag}(V). \tag{3.5}$$

Obviously, it holds $\hat{H}(\rho) = H(V(\rho))$. Therefore, X can be interpreted as an implicit function of V. Let $X(V) \in \mathbb{R}^{n\times p}$ be the eigenvectors corresponding to the p smallest eigenvalues of $H(V)$. Then, the fixed point map (3.3) is a system of nonlinear equations with respect to V as

$$V = \mathcal{V}(F_\phi(V)), \tag{3.6}$$

where $F_\phi(V) = \mathrm{diag}(X(V)X(V)^{\mathrm{T}})$.

The fixed point map (3.6) is well defined if there is a gap between the pth and $(p+1)$st smallest eigenvalues of $H(V)$. However, when these two eigenvalues are equal, there exists ambiguity on choosing the eigenvectors $X(V)$ since the multiplicity is greater than one. A common approach is to revise $F_\phi(V)$ in (3.6) by constructing a proper filter function. Let $q_1(V), \ldots, q_n(V)$ be the eigenvectors of $H(V)$ associated with eigenvalues $\lambda_1(V), \ldots, \lambda_n(V)$, respectively. A particular choice of the filter function is the Fermi–Dirac distribution of the form

$$f_\mu(t) := \frac{1}{1 + e^{\beta(t-\mu)}}, \tag{3.7}$$

where μ is the solution of the equations

$$\sum_{i=1}^{n} f_\mu(\lambda_i(V)) = p. \tag{3.8}$$

Since the left-hand side of (3.8) is monotonic with respect to μ for a fixed β, the solution to (3.8) is unique for any choice of β and λ_i. Then the fixed map (3.6) is replaced by the approximation

$$V = \mathcal{V}(F_{f_\mu}(V)), \tag{3.9}$$

where $F_{f_\mu}(V) = \operatorname{diag}(\sum_{i=1}^{n} f_\mu(\lambda_i(V)) q_i(V) q_i(V)^{\mathrm{T}})$ and μ satisfies (3.8).

3.2. The Jacobian of the fixed point maps. We first reformulate the functions $F_\phi(V)$ in (3.6) and $F_{f_\mu}(V)$ in (3.9) by spectral representation. Using the differentiability of spectral operators, they can be proved to be differentiable under some conditions. Let $\{\lambda_i(V), q_i(V)\}$ be the eigenpairs of $H(V)$ and assume that the eigenvalues $\lambda_1(V), \ldots, \lambda_n(V)$ are sorted in an ascending order, i.e., $\lambda_1(V) \le \cdots \le \lambda_p(V) \le \lambda_{p+1}(V) \le \cdots \le \lambda_n(V)$. The eigenvalue decomposition of $H(V)$ can be written as

$$H(V) = Q(V)\Pi(V)Q(V)^{\mathrm{T}}, \tag{3.10}$$

where $Q(V)$ and $\Pi(V)$ are

$$\begin{aligned} Q(V) &= [q_1(V),\ q_2(V),\ \ldots, q_n(V)] \in \mathbb{R}^{n\times n} \ \text{ and} \\ \Pi(V) &= \operatorname{Diag}(\lambda_1(V), \lambda_2(V), \ldots, \lambda_n(V)) \in \mathbb{R}^{n\times n}. \end{aligned} \tag{3.11}$$

Hence, the function $F_\phi(V)$ in (3.6) is equivalent to

$$F_\phi(V) = \ \operatorname{diag}(Q(V)\phi(\Pi(V))Q(V)^{\mathrm{T}}), \tag{3.12}$$

where $\phi(\Pi) = \operatorname{Diag}(\phi(\lambda_1(V)), \phi(\lambda_2(V)), \ldots, \phi(\lambda_n(V)))$ and

$$\phi(t) := \begin{cases} 1 & \text{for } t \le \frac{\lambda_p(V)+\lambda_{p+1}(V)}{2}, \\ 0 & \text{for } t > \frac{\lambda_p(V)+\lambda_{p+1}(V)}{2}. \end{cases} \tag{3.13}$$

Similarly, the function $F_{f_\mu}(V)$ in (3.9) in the spectral operator form is

$$F_{f_\mu}(V) = \operatorname{diag}(Q(V) f_\mu(\Pi(V)) Q(V)^{\mathrm{T}}). \tag{3.14}$$

Let $\mu_1, \ldots, \mu_{r(V)}$ be the distinct eigenvalues among $\{\lambda_1(V), \ldots, \lambda_n(V)\}$, $r(V)$ be the total number of distinct values, and $r_p(V)$ be the number of distinct eigenvalues

no greater than λ_p. For any $k = 1, \ldots, r(V)$, the set of indices i such that $\lambda_i = \mu_k$ is denoted by $\alpha_k := \{i \mid \lambda_i = \mu_k,\ i = 1, \ldots, n\}$. The next lemma shows the directional derivative of $F_\phi(V)$ by using the differentiability of the spectral operators [12, 15, 23, 29, 28].

LEMMA 3.1. *Suppose that Assumption* 2.5 *holds at $H(V)$, i.e., $\lambda_{p+1}(V) > \lambda_p(V)$. Then $F_\phi(V)$ is continuously differentiable and its directional derivative at V along $z \in R^n$ is*

$$\partial_V F_\phi(V)[z] = \operatorname{diag}\left(Q(V)\left(g_\phi(\Pi(V)) \circ \left(Q(V)^{\mathrm{T}} \operatorname{Diag}(z)\, Q(V)\right)\right) Q(V)^{\mathrm{T}}\right), \tag{3.15}$$

where "$\circ$" denotes the Hadamard product between two matrices, and $g_\phi(\Pi(V)) \in \mathbb{R}^{n\times n}$ is the so-called first divided difference matrix defined as

$$(g_\phi(\Pi(V)))_{ij} = \begin{cases} \frac{1}{\lambda_i(V)-\lambda_j(V)} & \text{if } i \in \alpha_k,\ j \in \alpha_l,\ k \le r_p(V),\ l > r_p(V), \\ \frac{-1}{\lambda_i(V)-\lambda_j(V)} & \text{if } i \in \alpha_k,\ j \in \alpha_l,\ k > r_p(V),\ l \le r_p(V), \\ 0 & \text{otherwise.} \end{cases} \tag{3.16}$$

Proof. The chain rule gives

$$\partial_V F_\phi(V)[z] = \frac{d \operatorname{diag}\left(Q\phi(\Pi)Q^{\mathrm{T}}\right)}{dH}\left[\partial_V H(V)[z]\right]. \tag{3.17}$$

By applying the continuous differentiability of the spectral operators in Proposition 2.10 of [15], the function $Q\phi(\Pi)Q^{\mathrm{T}}$ is differentiable with respect to H and its directional derivative is given by

$$\frac{dQ\phi(\Pi)Q^{\mathrm{T}}}{dH}[S] = Q\left(g_\phi(\Pi) \circ \left(Q^{\mathrm{T}}SQ\right)\right)Q^{\mathrm{T}} \quad \text{for all } S \in \mathbb{S}^n, \tag{3.18}$$

where, for any $i, j = 1, \ldots, n$,

$$(g_\phi(\Pi(V)))_{ij} = \begin{cases} \frac{\phi(\lambda_i(V))-\phi(\lambda_j(V))}{\lambda_i(V)-\lambda_j(V)} & \text{if } i \in \alpha_k,\ j \in \alpha_l,\ k \ne l, \\ 0 & \text{otherwise.} \end{cases} \tag{3.19}$$

Substituting (3.13) into (3.19) yields the specific form of $g_\phi(\pi(V))$ in (3.16). Since $\operatorname{diag}(\cdot)$ is a linear function, we obtain

$$\begin{aligned} \frac{d \operatorname{diag}\left(Q\phi(\Lambda)Q^{\mathrm{T}}\right)}{dH}[S] &= \frac{d \operatorname{diag}\left(Q\phi(\Lambda)Q^{\mathrm{T}}\right)}{dQ\phi(\Lambda)Q^{\mathrm{T}}} \frac{dQ\phi(\Lambda)Q^{\mathrm{T}}}{dH}[S] \\ &= \operatorname{diag}\left(Q\left(g_\phi(\Pi) \circ \left(Q^{\mathrm{T}}SQ\right)\right)Q^{\mathrm{T}}\right) \quad \text{for all } S \in \mathbb{S}^n. \end{aligned} \tag{3.20}$$

It follows from (3.5) that

$$\partial_V H(V)[z] = \operatorname{Diag}(z). \tag{3.21}$$

Plugging (3.20) and (3.21) into (3.17), we obtain (3.15). This completes the proof. ∎

We should point out that the Jacobian (3.15) is an alternative expression for the polarizability operator or linear response operator often used in the physics literature. It is the same as the well-known Adler–Wiser formula [1, 31].

Remark 3.2. Computing $\partial_V F_\phi(V)[z]$ requires all the eigenvectors $Q(V)$ and all eigenvalues $\Pi(V)$. Let $E_{j,p}$ ($O_{j,p}$) be the $j \times p$ matrix with ones (zeros) at all its entries. Then the matrix $g_\phi(\Pi(V)) \in \mathbb{R}^{n\times n}$ takes the specific form

$$g_\phi(\Pi(V)) = \begin{pmatrix} O_{p,p} & G \\ G^{\mathrm{T}} & O_{n-p,n-p} \end{pmatrix},$$

where

$$G = \begin{pmatrix} \frac{1}{\mu_1-\mu_{r_p(V)+1}} E_{|\alpha_1|,|\alpha_{r_p(V)+1}|} & \cdots & \frac{1}{\mu_1-\mu_{r(V)}} E_{|\alpha_1|,|\alpha_{r(V)}|} \\ \vdots & \ddots & \vdots \\ \frac{1}{\mu_{r_p(V)}-\mu_{r_p(V)+1}} E_{|\alpha_{r_p(V)}|,|\alpha_{r_p(V)+1}|} & \cdots & \frac{1}{\mu_{r_p(V)}-\mu_{r(V)}} E_{|\alpha_{r_p(V)}|,|\alpha_{r(V)}|} \end{pmatrix}.$$

The directional derivative of $F_{f_\mu}(V)[z]$ can be assembled in a similar fashion.

LEMMA 3.3. *The function $F_{f_\mu}(V)$ is continuously differentiable and its directional derivative at V along $z \in R^n$ is*

$$\partial_V F_{f_\mu}(V)[z] = \operatorname{diag}\left(Q(V)\left(g_{f_\mu}(\Pi(V)) \circ \left(Q(V)^{\mathrm{T}} \operatorname{Diag}(z)\, Q(V)\right)\right) Q(V)^{\mathrm{T}}\right), \tag{3.22}$$

where $g_{f_\mu}(\Pi(V)) \in \mathbb{R}^{n\times n}$ is defined as, for any $i, j = 1, \ldots, n$,

$$(g_{f_\mu}(\Pi(V)))_{ij} = \begin{cases} \frac{f_\mu(\lambda_i(V))-f_\mu(\lambda_j(V))}{\lambda_i(V)-\lambda_j(V)} & \text{if } i \in \alpha_k,\ j \in \alpha_l,\ k \neq l, \\ f'_\mu(\lambda_i(V)) & \text{otherwise.} \end{cases} \tag{3.23}$$

We next compute the Jacobian of $\mathcal{V}(F_\phi(V))$ and $\mathcal{V}(F_{f_\mu}(V))$.

THEOREM 3.4. *Let $J(\rho)$ be defined as* (2.3).

1. *Suppose that Assumption* 2.5 *holds at $H(V)$, i.e., $\lambda_{p+1}(V) > \lambda_p(V)$. Then the Jacobian of $\mathcal{V}(F_\phi(V))$ at V is*

$$\partial_V \mathcal{V}(F_\phi(V))[z] = J(F_\phi(V))\partial_V F_\phi(V)[z] \quad \text{for all } z \in \mathbb{R}^n. \tag{3.24}$$

2. *The Jacobian of $\mathcal{V}(F_{f_\mu}(V))$ at V is*

$$\partial_V \mathcal{V}(F_{f_\mu}(V))[z] = J(F_{f_\mu}(V))\partial_V F_{f_\mu}(V)[z] \quad \text{for all } z \in \mathbb{R}^n. \tag{3.25}$$

Proof. Note that

$$\partial_\rho(\mathcal{V}(\rho))[z] = J(\rho)z \quad \text{for all } z \in \mathbb{R}^n. \tag{3.26}$$

Applying the chain rule to $\partial_V \mathcal{V}(F_\phi(V))[z]$ and using (3.26) and (3.15), we obtain (3.24). This completes the proof. ∎

4. Convergence of the SCF iteration.

4.1. The SCF iteration and the simple mixing scheme. Starting from an initial vector $V^0 \in \mathbb{R}^n$, the SCF iteration for solving the fixed point map (3.6) recursively computes the eigenpairs $\{X(V^{i+1}), \Lambda(V^{i+1})\}$ as the solution of the linear eigenvalue problem

$$\begin{aligned} H(V^i)X(V^{i+1}) &= X(V^{i+1})\Lambda(V^{i+1}), \\ X(V^{i+1})^{\mathrm{T}}X(V^{i+1}) &= I, \end{aligned}$$

and then the potential is updated as

$$V^{i+1} = \mathcal{V}(F_\phi(V^i)). \tag{4.1}$$

When the difference between V^i and V^{i+1} is negligible, the system is said to be self-consistent and the SCF iteration is terminated.

The SCF iteration often converges slowly or even fails to converge. One of the heuristics for accelerating and stabilizing the SCF iteration is charge or potential mixing [18, 20]. Basically, the new potential V^{i+1} is constructed from a linear combination of the previously computed potential and the one obtained from certain schemes at the current iteration. In particular, the simple mixing scheme replaces (4.1) by updating

$$V^{i+1} = V^i - \alpha(V^i - \mathcal{V}(F_\phi(V^i))), \tag{4.2}$$

where α is a properly chosen step size. Similarly, the SCF iteration using simple mixing for solving the fixed point map (3.9) is

$$V^{i+1} = V^i - \alpha(V^i - \mathcal{V}(F_{f_\mu}(V^i))). \tag{4.3}$$

4.2. Global convergence analysis. We first make the following assumptions.

Assumption 4.1. The second-order derivatives of the exchange correlation functional $\epsilon_{xc}(\rho)$ are uniformly bounded from above. Without loss of generality, we assume that there exists a constant θ such that

$$\|J(\rho)\|_2 \le \theta \quad \text{for all } \rho \in \mathbb{R}^n. \tag{4.4}$$

Although we cannot verify Assumption 4.1 for any $X \in \mathbb{R}^{n\times p}$, it holds at a strong local minimizer using our lower bounds for nonzero charge densities in subsection 2.5 if the exchange correlation energy is (2.11).

It can be verified from the definition of the operator $\partial_V F_\phi(V)[\cdot]$ in (3.15) that it is a linear map. The induced ℓ_2-norms of $\partial_V \mathcal{V}(F_\phi(V))$ and $\partial_V F_\phi(V)[\cdot]$ are defined as

$$\|\partial_V \mathcal{V}(F_\phi(V))\|_2 = \max_{z\neq 0} \frac{\|\partial_V \mathcal{V}(F_\phi(V))[z]\|_2}{\|z\|_2} \text{ and } \|\partial_V F_\phi(V)\|_2 = \max_{z\neq 0} \frac{\|\partial_V F_\phi(V)[z]\|_2}{\|z\|_2}, \tag{4.5}$$

respectively. The next lemma shows that their ℓ_2-norms are bounded if Assumption 2.5 holds at $H(V)$.

LEMMA 4.2. *If Assumption* 2.5 *holds at $H(V)$ for a given $V \in \mathbb{R}^n$, then it holds*

$$\|\partial_V F_\phi(V)\|_2 \le \frac{1}{\delta} \quad \textit{and} \quad \|\partial_V \mathcal{V}(F_\phi(V))\|_2 \le \frac{\theta}{\delta}. \tag{4.6}$$

Proof. For any $z \in \mathbb{R}^n$, we obtain

$$\begin{aligned}
\|\partial_V F_\phi(V)[z]\|_2 &= \|\operatorname{diag}\left(Q(V)\left(g_\phi(\Pi(V)) \circ \left(Q(V)^{\mathrm{T}}\operatorname{Diag}(z)\,Q(V)\right)\right)Q(V)^{\mathrm{T}}\right)\|_2 \\
&\le \|Q(\rho)\left(g_\phi(\Pi(\rho)) \circ \left(Q(\rho)^{\mathrm{T}}\operatorname{Diag}(z)Q(\rho)\right)\right)Q(\rho)^{\mathrm{T}}\|_{\mathrm{F}} \\
&= \|g_\phi(\Pi(\rho)) \circ \left(Q(\rho)^{\mathrm{T}}\operatorname{Diag}(z)Q(\rho)\right)\|_{\mathrm{F}} \\
&\le \frac{1}{\delta}\|Q(\rho)^{\mathrm{T}}\operatorname{Diag}(z)Q(\rho)\|_{\mathrm{F}} \le \frac{1}{\delta}\|z\|_2,
\end{aligned} \tag{4.7}$$

where the second inequality is due to $|(g_\phi(\Pi(\rho)))_{ij}| \le 1/\delta$. Then the first inequality in (4.6) holds from the definitions (4.5) and (4.7). It follows from (3.24) and (4.7) that

$$\|\partial_V \mathcal{V}(F_\phi(V))[z]\|_2 \le \|J(F_\phi(V))\|_2 \|\partial_V F_\phi(V)[z]\|_2 \le \frac{\theta}{\delta}\|z\|_2. \tag{4.8}$$

This completes the proof. ∎

The set $\{H(V) \mid V \in \mathbb{R}^n\}$ is called uniformly well posed (UWP) [3, 32] with respect to a constant $\delta > 0$ if Assumption 2.5 holds at $H(V)$ with δ for any $V \in \mathbb{R}^n$. We next establish the convergence of the simple mixing scheme (4.2) when UWP holds.

THEOREM 4.3. *Suppose that Assumption* 4.1 *holds and* $\{H(V) \mid V \in \mathbb{R}^n\}$ *is UWP with a constant* δ *such that*

$$b_1 := 1 - \frac{\theta}{\delta} > 0. \tag{4.9}$$

Let $\{V^i\}$ *be a sequence generated by the simple mixing scheme* (4.2) *using a step size* α *satisfying*

$$0 < \alpha < \frac{2}{2 - b_1}. \tag{4.10}$$

Then $\{V^i\}$ *linearly converges to a solution of the KS equation* (1.5) *with linear convergence rate no worse than* $|1 - \alpha| + \alpha(1 - b_1)$.

Proof. For any V^i, it follows from (4.8), (4.9), and (4.10) that

$$\begin{aligned}\|(1-\alpha)I + \alpha\partial_V \mathcal{V}(F_\phi(V^i))\|_2 &\le |1-\alpha| + |\alpha|\|\partial_V \mathcal{V}(F_\phi(V^i))\|_2 \\ &\le \begin{cases} 1 - \alpha + \alpha\frac{\theta}{\delta} = 1 - \alpha b_1 & \text{if } 0 < \alpha < 1, \\ \alpha - 1 + \alpha\frac{\theta}{\delta} = \alpha(2 - b_1) - 1 & \text{if } \alpha \ge 1 \end{cases} \\ &< 1,\end{aligned}$$

which completes the proof. ∎

Remark 4.4. When the step size $\alpha = 1$, the simple mixing scheme (4.2) becomes the SCF iteration (4.1) with a convergence rate $\frac{\theta}{\delta}$. Since neither p nor n is involved in (4.9), it is much weaker than $\frac{12p\sqrt{n}\theta}{\delta} < 1$ for global convergence and $\frac{2\sqrt{n}\theta}{\delta} < 1$ for local convergence required by Theorems 3.3 and 4.2 in [25], respectively. In section 6, we numerically verify that condition (4.9) is satisfied on some practical examples.

We next establish convergence to the solutions of the modified fixed point map (3.9) without assuming the UWP properties.

THEOREM 4.5. *Suppose that Assumption* 4.1 *holds and*

$$b_2 := 1 - \frac{\beta\theta}{4} > 0. \tag{4.11}$$

Let $\{V^i\}$ *be a sequence generated by the simple mixing scheme* (4.3) *using a step size* α *satisfying* $0 < \alpha < \frac{2}{2-b_2}$. *Then the sequence* $\{V^i\}$ *converges to a solution of* (3.9) *with linear convergence rate no worse than* $|1 - \alpha| + \alpha(1 - b_2)$.

Proof. Using the mean value theorem and the fact that

$$|f'_\mu(t)| = \left|\frac{-\beta e^{\beta(t-\mu)}}{(1 + e^{\beta(t-\mu)})^2}\right| \le \frac{\beta}{4},$$

we obtain $|(g_{f_\mu}(\Pi(V)))_{ij}| \le \beta/4$, which yields $\|\partial_V \mathcal{V}(F_{f_\mu}(V))\|_2 \le \frac{\beta\theta}{4}$. Then, the convergence of (4.3) is proved similarly to that of Theorem 4.3. ∎

Remark 4.6. Suppose that UWP holds and f_μ is chosen such that

$$\begin{cases} \frac{1}{1+e^{\beta(\lambda_p - \mu)}} \ge 1 - \gamma, \\ \frac{1}{1+e^{\beta(\lambda_{p+1} - \mu)}} \le \gamma, \end{cases} \tag{4.12}$$

where $\gamma \ll 1$ is a constant. It can be shown that $\beta \geq \frac{2}{\delta} \cdot \ln \frac{1-\gamma}{\gamma}$. Hence, we have $\frac{\beta}{4} \geq \frac{1}{\delta}$ and the condition (4.9) is implied by (4.11) when $\ln \frac{1-\gamma}{\gamma} \geq 2$ or, equivalently, $\gamma \leq \frac{1}{e^2+1} \approx 0.12$. On the other hand, the closer γ is to zero, the closer f_μ is to ϕ from (4.12). Therefore, the convergence rate of the fixed point iteration using F_ϕ is better than that of F_{f_μ} when F_{f_μ} is sufficiently close to F_ϕ.

Remark 4.7. The convergence of the SCF iteration without simple mixing for solving a special KS equation without the exchange correlation energy is established in [32] under the condition $n^4\beta\|L^\dagger\|_2 < 2$. We can see that our condition is weaker since n^4 is not required.

Remark 4.8. It is necessary to make it clear that the solution to the modified problem (3.9) with a finite nonzero β is not the same as the solution to problem (3.9) with $\beta = \infty$ (zero temperature).

4.3. Local convergence analysis. Suppose that V^* is a solution of the fixed point map (4.2). Let $B(V^*, \eta) := \{V \mid ||V - V^*||_2 \leq \eta\}$ be a neighborhood of V^* for a given $\eta > 0$. The Taylor expansion at V^* yields

$$\begin{aligned} V^{k+1} - V^* &= V^k - \alpha(V^k - \mathcal{V}(F_\phi(V^k))) - (V^* - \alpha(V^* - \mathcal{V}(F_\phi(V^*)))) \\ &= (I - \alpha(I - \partial_V \mathcal{V}(F_\phi(V^*))))[V^k - V^*] + o(||V^k - V^*||_2) \\ &\qquad \text{for all } V^k \in B(V^*, \eta). \end{aligned} \tag{4.13}$$

If the spectral radius of the operator $I - \alpha(I - \partial_V \mathcal{V}(F_\phi(V^*)))$ is less than one, there must exist a sufficiently small η so that the simple mixing scheme (4.2) initiating from a point in $B(V^*, \eta)$ converges to V^* linearly.

We first present a few properties of the linear operators. Denote the space of linear operators by $\mathbb{L}(\mathbb{R}^n, \mathbb{R}^n) := \{\mathcal{P} \mid \mathcal{P} : \mathbb{R}^n \mapsto \mathbb{R}^n \text{ is a linear map}\}$. For a given $\mathcal{P} \in \mathbb{L}(\mathbb{R}^n, \mathbb{R}^n)$, if a scalar $\lambda \in \mathbb{C}$ and a nonzero vector $z \in \mathbb{C}^n$ satisfy $\mathcal{P}[z] = \lambda z$, the scalar λ and the vector z are called the eigenvalue and eigenvector of $\mathcal{P}$, respectively. Then the spectral radius, denoted by $\varrho(\mathcal{P})$, is the largest absolute value of eigenvalues of $\mathcal{P}$. A linear operator $\mathcal{P}$ is called symmetric if $y^{\mathrm{T}}\mathcal{P}[x] = x^{\mathrm{T}}\mathcal{P}[y]$ for any $x, y \in \mathbb{R}^n$, and is called positive semidefinite if $z^{\mathrm{T}}\mathcal{P}[z] \geq 0$ for all $z \in \mathbb{R}^n$.

The next lemma shows that $\partial_V F_\phi(V)[\cdot]$, as a linear operator, is negative semidefinite.

LEMMA 4.9. *For any $z \in \mathbb{R}^n$, it holds $z^{\mathrm{T}}\partial_V F_\phi(V)[z] \leq 0$.*

Proof. For any $z \in R^n$, we have

$$\begin{aligned} z^{\mathrm{T}}\partial_V F_\phi(V)[z] &= z^{\mathrm{T}} \operatorname{diag}\left(Q(V)\left(g_\phi(\Pi(V)) \circ \left(Q(V)^{\mathrm{T}}\operatorname{Diag}(z)\, Q(V)\right)\right) Q(V)^{\mathrm{T}}\right) \\ &= \left\langle \left(Q(V)^{\mathrm{T}}\operatorname{Diag}(z)\, Q(V)\right), g_\phi(\Pi(V)) \circ \left(Q(V)^{\mathrm{T}}\operatorname{Diag}(z)\, Q(V)\right)\right\rangle \\ &= e^{\mathrm{T}}\left(g_\phi(\Pi(V)) \circ \left(Q(V)^{\mathrm{T}}\operatorname{Diag}(z)\, Q(V)\right) \circ \left(Q(V)^{\mathrm{T}}\operatorname{Diag}(z)\, Q(V)\right)\right) e \\ &\leq 0, \end{aligned}$$

where the third equality uses the properties of the Hadamard products and the inequality is due to

$$\left(Q(V)^{\mathrm{T}}\operatorname{Diag}(z)\, Q(V)\right) \circ \left(Q(V)^{\mathrm{T}}\operatorname{Diag}(z)\, Q(V)\right) \geq 0 \text{ and } g_\phi(\Pi(V)) \leq 0.$$

This completes the proof. ∎

The next lemma shows that the eigenvalues of the product of a symmetric matrix and a symmetric positive semidefinite linear operator are real.

LEMMA 4.10. *Suppose that $M \in \mathbb{R}^{n\times n}$ is a symmetric matrix and $\mathcal{P} \in \mathbb{L}(\mathbb{R}^n, \mathbb{R}^n)$ is a symmetric positive semidefinite linear operator. Then all the eigenvalues of the linear operator $M\mathcal{P}$ are real. Furthermore, it holds*

$$\lambda_{\max}(M\mathcal{P}) \le \begin{cases} \lambda_{\max}(M)\lambda_{\max}(\mathcal{P}) & \text{if } \lambda_{\max}(M) \ge 0, \\ \lambda_{\max}(M)\lambda_{\min}(\mathcal{P}) & \text{otherwise}, \end{cases} \tag{4.14}$$

$$\lambda_{\min}(M\mathcal{P}) \ge \begin{cases} \lambda_{\min}(M)\lambda_{\min}(\mathcal{P}) & \text{if } \lambda_{\min}(M) \ge 0, \\ \lambda_{\min}(M)\lambda_{\max}(\mathcal{P}) & \text{otherwise}. \end{cases} \tag{4.15}$$

Proof. The proof is omitted since it follows by standard arguments. □

We now establish the local convergence result for the simple mixing scheme.

THEOREM 4.11. *Let V^* be a solution of the KS equation* (1.5)*. Suppose that Assumption* 4.1 *holds and Assumption* 2.5 *is valid at $H(V^*)$ with a constant δ satisfying*

$$\delta > -\lambda^*_{\min}, \tag{4.16}$$

where $\lambda^*_{\min} := \min\{0, \lambda_{\min}(J(F_\phi(V^*)))\}$. There exists an open neighborhood Ω of V^*, such that the sequence $\{V^i\}$ generated by the simple mixing scheme (4.2) *using $V^0 \in \Omega$ and a step size*

$$\alpha \in \left(0, \frac{2\delta}{\theta+\delta}\right) \tag{4.17}$$

converges to V^* with linear convergence rate no worse than

$$\max\{|1-\alpha\frac{\delta+\lambda^*_{\min}}{\delta}|, |\alpha\frac{\theta+\delta}{\delta}-1|\}.$$

Proof. The Taylor expansion (4.13) implies that local convergence of the scheme (4.2) holds if

$$\varrho(I-\alpha\mathcal{A}) < 1, \tag{4.18}$$

where $\mathcal{A} := I - J(F_\phi(V^*))\partial_V F_\phi(V^*)$. According to Lemma 4.9, $-\partial_V F_\phi(V^*)$ is symmetric positive semidefinite. Using Lemma 4.10, we conclude that all the eigenvalues of $\mathcal{A}$ are real. Hence, (4.18) is guaranteed if

$$\lambda_{\min}(\mathcal{A}) > 0, \tag{4.19}$$

$$\alpha\lambda_{\max}(\mathcal{A}) < 2. \tag{4.20}$$

Note that $\lambda_{\min}(\mathcal{A}) = 1 + \lambda_{\min}(J(F_\phi(V^*))(-\partial_V F_\phi(V^*)))$. Using Lemma 4.10, $\lambda_{\max}(-\partial_V F_\phi(V^*)) \le \frac{1}{\delta}$ from Lemma 4.2, and the definition of $\lambda^*_{\min}$, we obtain

$$\begin{aligned} \lambda_{\min}(\mathcal{A}) - 1 &\ge \begin{cases} \lambda_{\min}(J(F_\phi(V^*)))\lambda_{\min}(-\partial_V F_\phi(V^*)) & \text{if } \lambda_{\min}(J(F_\phi(V^*))) \ge 0, \\ \lambda_{\min}(J(F_\phi(V^*)))\lambda_{\max}(-\partial_V F_\phi(V^*)) & \text{otherwise} \end{cases} \\ &\ge \begin{cases} 0 & \text{if } \lambda_{\min}(J(F_\phi(V^*))) \ge 0, \\ \frac{1}{\delta}\lambda_{\min}(J(F_\phi(V^*))) & \text{otherwise} \end{cases} \\ &\ge \frac{\lambda^*_{\min}}{\delta}, \end{aligned}$$

which yields (4.19) from the assumption (4.16).

Using Lemma 4.10 again, we have

$$\begin{aligned}\lambda_{\max}(\mathcal{A}) &\le 1+\lambda_{\max}(J(F_\phi(V^*))(-\partial_V F_\phi(V^*)))\\ &\le 1+\max\{0,\lambda_{\max}(J(F_\phi(V^*)))\lambda_{\max}(-\partial_V F_\phi(V^*))\}\le 1+\frac{\theta}{\delta},\end{aligned}\tag{4.21}$$

which together with (4.17) gives (4.20). ∎

COROLLARY 4.12. *Suppose that Assumption* 2.5 *holds at* $H(V^*)$ *and* $J(F_\phi(V^*))$ *is positive semidefinite. Then the condition* (4.16) *holds.*

Remark 4.13. Suppose that Assumption 2.5 holds at $H(V^*)$. The condition $\delta > -\lambda^*_{\min}$ is much weaker than the condition $\delta > \theta$ for ensuring global convergence required by Theorem 4.3. In fact, it is clear that the condition (4.16) holds when $J(F_\phi(V^*))$ is positive semidefinite. Since $J((F_\phi(V^*))) = L^\dagger + \frac{\partial \mu_{xc}(F_\phi(V^*))}{\partial F_\phi(V^*)} e$ and $L^\dagger$ is positive semidefinite, the second-order derivative of the exchange correlation energy plays a more important role when $\lambda_{\min}(J((F_\phi(V^*)))) < 0$.

We can obtain the following local convergence result for the modified fixed point map (3.9) in the same manner as Theorem 4.5.

COROLLARY 4.14. *Suppose that Assumption* 4.1 *holds and* $\frac{4}{\beta} > -\lambda^*_{\min}$, *where* $\lambda^*_{\min} := \min\{0, \lambda_{\min}(J(F_{f_\mu}(V^*)))\}$. *Let* V^* *be a solution of the KS equation* (1.5). *There exists an open neighborhood* Ω *of* V^* *such that the sequence* $\{V^i\}$ *generated by the simple mixing scheme* (4.3) *using* $V^0 \in \Omega$ *and a step size*

$$\alpha \in \left(0, \frac{8}{\theta\beta + 4}\right)\tag{4.22}$$

converges to V^* *with linear convergence rate no worse than*

$$\max\{|1-\alpha(\frac{\lambda^*_{\min}\beta+4}{4})|, |\alpha(\frac{\theta\beta+4}{4})-1|\}.$$

5. Convergence analysis of approximate Newton approaches. The generalized Jacobian $\partial_V \mathcal{V}(F(V))$ in (3.24) suggests that Newton's method for solving the fixed point map (3.6) is

$$V^{i+1} = V^i - \alpha\left(I - J(F_\phi(V^i))\partial_V F_\phi(V^i)\right)^{-1}\left(V^i - \mathcal{V}\left(F_\phi(V^i)\right)\right),$$

where α is a step size. Obviously, this method is not computationally practical for solving the fixed point maps due to the presence of all eigenvectors and eigenvalues in $\partial_V F_\phi(V)[\cdot]$. In this section, we propose two approximate Newton approaches in the form

$$V^{i+1} = V^i - \alpha\left(I - D^i\right)^{-1}\left(V^i - \mathcal{V}\left(F_\phi(V^i)\right)\right),\tag{5.1}$$

where $\alpha > 0$ and $D^i \in \mathbb{R}^{n\times n}$ is a matrix for approximating the Jacobian $\partial_V \mathcal{V}(F(V^i))$.

THEOREM 5.1. *Suppose that Assumption* 4.1 *and UWP hold. Let* $\{V^i\}$ *be a sequence generated by* (5.1) *using* $\{D^i\}$ *and a step size* α *such that*

$$0 < \alpha < \frac{2}{b_2}, \quad 0 < \gamma_{\min} := \sigma_{\min}(I - D^i), \text{ and } \gamma_{\max} := \sigma_{\max}(I - D^i),$$

where $b_2 := 1 + \frac{\theta}{\delta}$, *and* $\sigma_{\min}$ *and* $\sigma_{\max}$ *are the smallest and largest singular values of* $I - D^i$, *respectively. If* $b_1 := 1 - \frac{\gamma_{\max}}{\gamma_{\min}}\frac{\theta}{\delta} > 0$, *then* $\{V^i\}$ *converges to a solution of the KS equation* (1.5) *with linear convergence rate no more than*

$$\max(|1-\alpha\gamma^{-1}_{\max}b_1|, |\alpha\gamma^{-1}_{\min}b_2 - 1|).$$

Proof. For any V^i, it follows from the definitions of D^i, α, and b_2 that

$$\begin{aligned}
&\|I-\alpha(I-D^i)^{-1}(I-\partial_V\mathcal{V}(F_\phi(V^i)))\|_2\\
&\quad=\|I-\alpha(I-D^i)+\alpha(I-D^i)^{-1}\partial_V\mathcal{V}(F_\phi(V^i))\|_2\\
&\quad\le\|I-\alpha(I-D^i)\|_2+|\alpha|\|(I-D^i)^{-1}J(F_\phi(V^i))J(V^i)\|_2\\
&\quad\le\begin{cases}1-\alpha\gamma_{\max}^{-1}+\alpha\gamma_{\min}^{-1}\frac{\theta}{\delta}=1-\alpha\gamma_{\max}^{-1}b_1 & \text{if}\,\alpha<\gamma_{\max};\\ \alpha\gamma_{\min}^{-1}-1+\alpha\gamma_{\min}^{-1}\frac{\theta}{\delta}=\alpha\gamma_{\min}^{-1}b_2-1 & \text{otherwise}\end{cases}\\
&\quad<1.
\end{aligned}$$

This completes the proof. ∎

Remark 5.2. Although $\delta>\theta$ is also required for the global convergence of the approximate Newton method similarly to Theorem 4.3 and the theoretical choice of D^i should satisfy $\frac{\gamma_{\max}}{\gamma_{\min}}<\delta/\theta$, these conditions may be relaxed in practical computation.

5.1. Approximate Newton method I. Our first approach replaces the operator $\partial_V F_\phi(V^i)[\cdot]$ by a diagonal matrix $\tau^i I$, where τ^i is a nonpositive scalar. It is chosen to be nonpositive since $\partial_V F_\phi(V^i)[\cdot]$ is negative semidefinite from Lemma 4.9. Consequently, we set $D^i:=\tau^i J(\rho)$ and the scheme (5.1) becomes

$$V^{i+1}=V^i-\alpha\left(I-\tau^i J(F_\phi(V^i))\right)^{-1}\left(V^i-\mathcal{V}\left(F_\phi(V^i)\right)\right). \tag{5.2}$$

The next theorem presents the local convergence analysis for the method (5.2).

THEOREM 5.3. *Let V^* be a solution of the KS equation* (1.5)*. Suppose that Assumption* 4.1 *holds with a constant θ and Assumption* 2.5 *is valid at $H(V^*)$ with a constant δ satisfying*

$$\frac{\delta^2}{2\delta+\theta}>-\lambda_{\min}^*, \tag{5.3}$$

where $\lambda_{\min}^*:=\min\{0,\lambda_{\min}(J(F_\phi(V^*)))\}$. Let $\{V^i\}$ be a sequence generated by the scheme (5.2) *using $\lim_{i\to\infty}\tau^i=\tau^*\in(-\frac{1}{\delta},0)$ and a step size*

$$\alpha\in\left(0,\frac{2(\delta+\lambda_{\min}^*)}{\theta+\delta}\right). \tag{5.4}$$

If the initial point V^0 is selected in a sufficiently small open neighborhood of V^*, then $\{V^i\}$ converges to V^* with linear convergence rate no worse than

$$\max\{|1-\alpha(\frac{\delta}{\theta+\delta}+\frac{\lambda_{\min}^*}{\delta+\lambda_{\min}^*})|,|\alpha(\frac{\theta+\delta}{\delta+\lambda_{\min}^*})-1|\}.$$

Proof. The convergence of the iteration (5.2) is guaranteed by

$$\varrho(I-\alpha\mathcal{M})<1, \tag{5.5}$$

where $\mathcal{M}=\mathcal{M}_1^{-1}(I-J(F_\phi(V^*))\partial_V F_\phi(V^*))$ and $\mathcal{M}_1=I-\tau^* J(F_\phi(V^*))$. A direct linear algebraic calculation yields $\mathcal{M}=I+\mathcal{M}_1^{-1}J(F_\phi(V^*))(\tau^* I-\partial_V F_\phi(V^*))$. By Lemma 4.10, the symmetry of $J(F_\phi(V^*))$, and the symmetric positiveness of $\tau^* I-\partial_V F_\phi(V^*)$, we see that all eigenvalues of $\mathcal{M}$ are real. Then in order to obtain (5.5), it suffices to have $\lambda_{\min}(\mathcal{M})>0$ and $\alpha\lambda_{\max}(\mathcal{M})<2$. Using $0>\tau^*>-\frac{1}{\delta}$ and the definition of $\lambda_{\min}^*$, we have

$$\lambda_{\min}(\mathcal{M}_1)\ge\frac{\delta+\lambda_{\min}^*}{\delta}>0\ \text{and}\ \lambda_{\max}(\mathcal{M}_1)\le\frac{\theta+\delta}{\delta}. \tag{5.6}$$

Using the fact that the smallest eigenvalue of a summation of two matrices is larger than the summation of the smallest eigenvalues of these matrices, we obtain

$$\begin{aligned}\lambda_{\min}(\mathcal{M}) &\geq \lambda_{\min}(\mathcal{M}_1^{-1}) + \lambda_{\min}(\mathcal{M}_1^{-1}J(F_\phi(V^*))(-\partial_V F_\phi(V^*))) \\ &\geq \frac{\delta}{\theta+\delta} + \lambda_{\min}(\mathcal{M}_1^{-1}J(F_\phi(V^*))(-\partial_V F_\phi(V^*))).\end{aligned} \tag{5.7}$$

Applying Lemma 4.10, $\lambda_{\max}(-\partial_V F_\phi(V^*)) \leq \frac{1}{\delta}$ from Lemma 4.2, and the definition of $\lambda^*_{\min}$, we have

$$\begin{aligned}&\lambda_{\min}(\mathcal{M}_1^{-1}J(F_\phi(V^*))(-\partial_V F_\phi(V^*))) \\ &\quad\geq \begin{cases}\lambda_{\min}(\mathcal{M}_1^{-1})\lambda_{\min}(J(F_\phi(V^*)))\lambda_{\min}(-\partial_V F_\phi(V^*)) & \text{if } \lambda_{\min}(J(F_\phi(V^*))) \geq 0, \\ \lambda_{\max}(\mathcal{M}_1^{-1})\lambda_{\min}(J(F_\phi(V^*)))\lambda_{\max}(-\partial_V F_\phi(V^*)) & \text{otherwise}\end{cases} \\ &\quad\geq \begin{cases}0 & \text{if } \lambda_{\min}(J(F_\phi(V^*))) \geq 0, \\ \frac{\lambda_{\min}(J(F_\phi(V^*)))}{\delta+\lambda^*_{\min}} & \text{otherwise}\end{cases} \\ &\quad\geq \frac{\lambda^*_{\min}}{\delta+\lambda^*_{\min}},\end{aligned}$$

which together with (5.7) gives $\lambda_{\min}(\mathcal{M}) > 0$. It follows from Lemma 4.10 and (5.6) that

$$\begin{aligned}\lambda_{\max}(\mathcal{M}) &\leq \lambda_{\max}(\mathcal{M}_1^{-1}) + \lambda_{\max}(\mathcal{M}_1^{-1})\lambda_{\max}(J(F_\phi(V^*)))\lambda_{\max}(-\partial_V F_\phi(V^*)) \\ &\leq \frac{\theta+\delta}{\delta+\lambda^*_{\min}}.\end{aligned} \tag{5.8}$$

Combining (5.4) and (5.8) yields $\alpha\lambda_{\max}(\mathcal{M}) < 2$. ∎

Remark 5.4. The condition (5.3) implies that $\delta > -2\lambda^*_{\min}$. Hence, it is much weaker than the condition required by global convergence in Theorem 5.1. Similarly to the simple mixing, the condition (5.3) holds when $J(F_\phi(V^*))$ is positive semidefinite.

5.2. Approximate Newton method II. The matrix $J(\rho)$ has to be calculated for each ρ in the approximate Newton method (5.2). If the computational cost of second-order derivatives of the exchange correlation function is expensive, a simpler choice is to approximate $J(F_\phi(V^*))$ by $L^\dagger$ and $\partial_V F_\phi(V)$ by $\tau^i I$, that is, $D^i = \tau^i L^\dagger$. Hence, the approximate Newton method (5.1) becomes

$$V^{i+1} = V^i - \alpha\left(I - \tau^i L^\dagger\right)^{-1}\left(V^i - \mathcal{V}\left(F_\phi(V^i)\right)\right), \tag{5.9}$$

where $\{\tau^i\}$ is negative. In fact, (5.9) is the Kerker preconditioner [18, 24].

THEOREM 5.5. *Let V^* be a solution of the KS equation* (1.5)*. Suppose that Assumption* 4.1 *holds with a constant θ and Assumption* 2.5 *is valid at $H(V^*)$ with a constant δ satisfying $\delta > \theta$. Let $\{V^i\}$ be a sequence generated by the scheme* (5.2) *using $\lim_{i\to\infty}\tau_i = \tau^* \in (-\frac{1}{\xi}, 0)$ such that $\xi \geq \frac{||L^\dagger||_2\theta}{\delta-\theta}$, and a step size*

$$\alpha \in \left(0, \frac{2}{1+\frac{\theta}{\delta}}\right). \tag{5.10}$$

If the initial point V^0 is selected in a sufficiently small open neighborhood of V^*, then $\{V^i\}$ converges to V^* with linear convergence rate no worse than

$$\max\{|1-\alpha(\frac{\xi}{||L^\dagger||_2+\xi} - \frac{\theta}{\delta})|, |\alpha(1+\frac{\theta}{\delta})-1|\}$$

Proof. Let $\bar{\mathcal{M}} = (I-\tau^* L^\dagger)^{-1}(I-\partial_V \mathcal{V}(F_\phi(V^*)))$. The convergence of the iteration (5.9) is guaranteed by

$$\varrho(I - \alpha\bar{\mathcal{M}}) < 1. \tag{5.11}$$

We can decompose $\bar{\mathcal{M}} = \bar{\mathcal{M}}_1 - \bar{\mathcal{M}}_2$, where $\bar{\mathcal{M}}_1 = (I - \tau^* L^\dagger)^{-1}$ and $\bar{\mathcal{M}}_2 = (I - \tau^* L^\dagger)\partial_V \mathcal{V}(F_\phi(V^*))$. Since $L^\dagger$ is positive semidefinite and $\tau^* \le 0$, we have

$$\lambda_{\min}(\bar{\mathcal{M}}_1) > \frac{\xi}{||L^\dagger||_2 + \xi}, \tag{5.12}$$

$$\lambda_{\max}(\bar{\mathcal{M}}_1) \le 1. \tag{5.13}$$

Using Assumption 4.1 and Lemma 4.2, we have

$$||\bar{\mathcal{M}}_2||_2 = ||(I - \tau^* L^\dagger)^{-1}\partial_V \mathcal{V}(F_\phi(V^*))||_2 \le \frac{\theta}{\delta}. \tag{5.14}$$

Using (5.12) and $\xi \ge \frac{||L^\dagger||_2\theta}{\delta-\theta}$, we obtain $\lambda_{\min}(\bar{\mathcal{M}}_1) > \frac{\theta}{\delta}$, which together with (5.14) yields

$$(1 - \alpha\lambda_{\min}(\bar{\mathcal{M}}_1)) < 1 - \alpha||\bar{\mathcal{M}}_2||_2. \tag{5.15}$$

On the other hand, it follows from (5.10), (5.13), and (5.14) that

$$(\alpha\lambda_{\max}(\bar{\mathcal{M}}_1) - 1) < 1 - \alpha||\bar{\mathcal{M}}_2||_2. \tag{5.16}$$

Combining (5.15) and (5.16) gives $\varrho(1 - \alpha\bar{\mathcal{M}}_1) < 1 - \alpha||\bar{\mathcal{M}}_2||_2$. which guarantees (5.11). ∎

Remark 5.6. Although Theorems 5.3 and 5.5 can hardly be considered satisfactory, they might be valuable for understanding the behaviors of these approximate Newton approaches.

6. Numerical results. In this section, we verify the assumptions of our global and local convergence on a few simple examples in a MATLAB toolbox KSSOLV [33]. The continuous problem is discretized by using the planewave discretization scheme. Throughout what follows, we adhere to the convention used in the implementation of the KSSOLV toolbox. Roughly speaking, the variable X is represented as $[F(\hat{X}_1), \ldots, F(\hat{X}_p)]$, where each $\hat{X}_i$ corresponds to a 3D matrix of size $n_1 \times n_2 \times n_3$ and F is the 3D discretized Fourier transform operator such that $F^*F = NI$, $FF^* = NI$, and $N = n_1 n_2 n_3$. Let $g_j \in \mathbb{R}^3$ be frequency vectors arranged in a lexicographical order and D_g be a diagonal matrix whose diagonal entries d_j are

$$d_j = \begin{cases} \|g_j\|^2 & \text{if } \|g_j\|^2 < 2E_{cut}, \\ 0 & \text{otherwise,} \end{cases}$$

where E_{cut} is the kinetic energy cutoff value. Then the Laplacian operator can be discretized as $L = \frac{1}{N}F^* D_g F$ and the operator $L^\dagger$ is defined as $\frac{1}{N}F^* D_g^\dagger F$. Furthermore, KSSOLV only saves the components of $\hat{X}_i$ whose indices correspond to the nonzero elements of d_j. The number of nonzero elements of d_j is denoted as the dimension n. SCF in KSSOLV is performed with respect to $\hat{X}$ and the corresponding Hamiltonian matrix is

$$\hat{H}(V) := \frac{1}{2}D_g + \frac{1}{N}F\left(V_{ion} + \mathrm{Diag}(V)\right)F^*. \tag{6.1}$$

Clearly, the eigenvalue $\lambda_i(H(V))$, where $H(V)$ is defined in (1.4), is equal to $N\lambda_i(\hat{H}(V))$.

TABLE 1
Numerical results of the SCF using the simple mixing scheme (4.2) *with* $\alpha = 0.5$.

name	(n_1, n_2, n_3)	n	p	iter	res	d_j	$\left\|\text{diag}\left(\frac{\partial \mu_{xc}(\rho)}{\partial \rho} e\right)\right\|$	ratio1	ratio2
c2h6	(32, 32, 32)	2103	7	17	1.14e-06	(1.97e-01, 4.99e+01)	(5.97e-01, 1.11e+04)	1.2860	1.2860
co2	(32, 32, 32)	2103	8	22	9.98e-08	(1.97e-01, 4.99e+01)	(3.83e-01, 1.85e+04)	2.7901	2.7901
h2o	(32, 32, 32)	2103	4	17	1.31e-06	(1.97e-01, 4.99e+01)	(4.02e-01, 1.87e+05)	33.7847	33.7847
hnco	(32, 32, 32)	2103	8	22	1.81e-07	(1.97e-01, 4.99e+01)	(3.62e-01, 1.05e+04)	4.1224	4.1223
nic	(16, 16, 16)	251	7	20	2.03e-07	(7.90e-01, 4.90e+01)	(5.21e-01, 4.26e+00)	0.0067	0.0047
si2h4	(32, 32, 32)	2103	6	18	1.69e-06	(1.97e-01, 4.99e+01)	(5.60e-01, 4.67e+03)	0.9864	0.9864
sih4	(32, 32, 32)	2103	4	16	1.12e-06	(1.97e-01, 4.99e+01)	(8.92e-01, 2.32e+03)	0.2954	0.2954

The exchange-correlation potential is computed by the Perdew–Zunger formula. Let $\tilde{\gamma} = e^2(\frac{3}{\pi})^{1/3}$ and $\phi_i = (4\pi\rho_i/3)^{-1/3}$. The function ϵ_{xc} in (1.2) is defined by $(\epsilon_{xc})_i = ((\epsilon_{ex})_i + (\epsilon_{ec})_i)\rho_i$, where $(\epsilon_{ex})_i = -\frac{3}{4}\tilde{\gamma}\rho_i^{\frac{1}{3}}$ and

$$(\epsilon_{ec})_i = \begin{cases} A_1 + A_2\phi_i + (A_3 + A_4\phi_i)\log(\phi_i) & \text{if } \phi_i < 1, \\ \frac{B_1}{1+B_2\sqrt{\phi_i}+B_3\phi_i} & \text{if } \phi_i \geq 1, \end{cases}$$

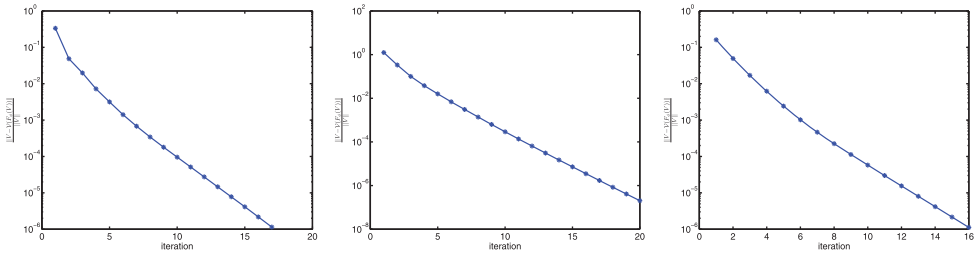
where $A_1 = -0.096, A_2 = -0.0232, A_3 = 0.0622, A_4 = 0.004, B_1 = -0.2846, B_2 = 1.0529$, and $B_3 = 0.3334$. The matrix $\frac{\partial \mu_{xc}(\rho)}{\partial \rho} e$ in $J(F_\phi(V))$ is a diagonal matrix in this case.

We apply the SCF using the simple mixing scheme (4.2) with $\alpha = 0.5$ to solve seven problems including c2h6, co2, h2o, hnco, nic, si2h4, and sih4. The value "E_{cut}" is set to 25 in all examples. A summary of numerical results is presented in Table 1. In the table, "iter" denotes the number of iterations performed by the SCF, and "res," "ratio1," and "ratio2" are the values of $\frac{\|V-\mathcal{V}(F_\phi(V))\|}{\|V\|}$, $\frac{\|J(F_\phi(V))\|_2}{\delta}$, and $\frac{-\min\{0,\lambda_{\min}(J(F_\phi(V)))\}}{\delta}$ computed at the final iteration, respectively. The smallest and largest values of the nonzero elements of d_j and $|\,\text{diag}(\frac{\partial \mu_{xc}(\rho)}{\partial \rho} e)|$ at the final iteration are also reported. From this table, we can see that the SCF solves all problems successfully to an accuracy of order 10^{-6}. For a further illustration of the convergence properties, the values of $\frac{\|V-\mathcal{V}(F_\phi(V))\|}{\|V\|}$ and $\frac{\|J(F_\phi(V)))\|_2}{\lambda_{p+1}(H(V))-\lambda_p(H(V))}$ at each iteration of c2h6, nic, and sih4 are depicted in Figures 1 and 2, respectively.

Our gap assumption holds on nic, si2h4, and sih4. Although it fails on c2h6, co2, h2o, and hnco, this assumption might be further improved given the good performance of the SCF on these examples and because our estimation $\|\partial_V F_\phi(V)\|_2 \leq \frac{1}{\delta}$ in Lemma 4.2 is not tight.

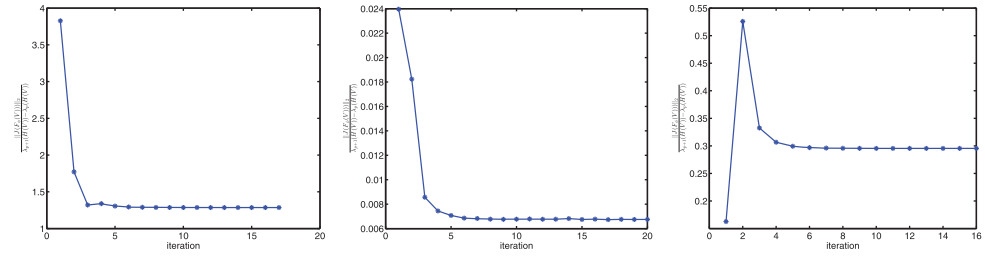
Remark 6.1. Our current conditions may not hold for large systems. As one referee pointed out, the primitive supercell on which the computations take place can be quite large in many scenarios. This is the case, for instance, for a simulation box of liquid containing many atoms, or for a complex crystal, or even a simple crystal under nonideal situations. Then the norm of $L^\dagger$ can become quite large and our assumptions on the gap δ can hardly be realized since $J(\rho) := L^\dagger + \frac{\partial \mu_{xc}(\rho)}{\partial \rho} e$.

7. Conclusion. The equivalence between the KS total energy minimization problem and the KS equation is ambiguous in the current literatures on KS DFT. A simple counterexample shows that the solutions of these two problems are not necessarily the same. We examine the equivalence based on the optimality conditions for a specialized exchange correlation functional. We prove that a global solution of the KS minimization problem is a solution of the KS equation if the gap between the pth



(a) c2h6 (b) nic (c) sih4

FIG. 1. *Numerical results on the convergence of* $\frac{\|V-\mathcal{V}(F_\phi(V))\|}{\|V\|}$.



(a) c2h6 (b) nic (c) sih4

FIG. 2. *Numerical results on the convergence of* $\frac{\|J(F_\phi(V)))\|_2}{\lambda_{p+1}(H(V))-\lambda_p(H(V))}$.

TABLE 2

Equivalence between the KS total energy minimization and the KS equation using the exchange correlation function $e^{\mathrm{T}}\epsilon_{xc}(\rho) = -\frac{3}{4}\gamma\rho^{\mathrm{T}}\rho^{\frac{1}{3}}$.

Properties	Eigenvalue gap δ	Other assumptions
A global minimizer X^* solves the KS equation	Assumption 2.5 holds at $H(X^*)$ with $\delta > p\left(\|\|L^\dagger\|\|_2 - \frac{\gamma}{3}\right)$	–
A local minimizer X^* solves the KS equation	Assumption 2.5 holds at $H(X^*)$ with $\delta > 2\left(\|\|L^\dagger\|\|_2 - \frac{\gamma}{3}\right)$	$\rho_i > 0$, $i = 1, \ldots, n$
$\rho_i(X^*) \in [0, c) \Rightarrow \rho_i(X^*) = 0$	–	X^* is a strong local minimizer

and $(p+1)$st smallest eigenvalues of the Hamiltonian $H(X)$ is sufficiently large. The equivalence of a local minimizer requires that the corresponding charge densities are all positive. For strong local minimizers, the nonzero charge densities are bounded below by a positive constant uniformly. These properties are summarized in Table 2.

We improve the convergence analysis on the SCF iteration for solving the KS equation by analyzing the Jacobian of the corresponding fixed point maps. Global convergence of the simple mixing scheme can be established when there exists a gap between the pth and $(p+1)$st eigenvalues of the Hamiltonian $H(X)$. This assumption can be relaxed for local convergence analysis and if the charge density is computed using the Fermi–Dirac distribution. Our results require much weaker conditions than the previous analysis in [25]. The structure of the Jacobian also suggests two approximate Newton methods. In particular, the second one is exactly the method of Kerker preconditioner proposed in [24]. Although our assumption on the gap is very stringent and is almost never satisfied in reality, our analysis is helpful for a better un-

TABLE 3
Convergence results for solving the KS equation under Assumption 4.1.

Properties		Eigenvalue gap δ or smoothing parameter β	Step size α
Analysis of SCF in [25]	global convergence	UWP holds and $\delta > 12p\sqrt{n}\theta$	1
	local convergence	Assumption 2.5 holds at the local minimizer with $\delta > 2\sqrt{n}\theta$	1
Analysis of SCF with simple mixing	global convergence using F_ϕ	UWP holds with $\delta > \theta$	$\left(0, \frac{2\delta}{\theta+\delta}\right)$
	global convergence using F_{f_μ}	$\frac{4}{\beta} > \theta$	$\left(0, \frac{8}{\theta\beta+4}\right)$
	local convergence using F_ϕ	Assumption 2.5 holds at the local minimizer with $\delta > -\min\{0, \lambda_{\min}(J(F_\phi(V^*)))\}$	$\left(0, \frac{2\delta}{\theta+\delta}\right)$
	local convergence using F_{f_μ}	$\frac{4}{\beta} > -\min\{0, \lambda_{\min}(J(F_\phi(V^*)))\}$	$\left(0, \frac{8}{\theta\beta+4}\right)$
Analysis of Approximate Newton methods	global convergence	UWP holds with $\delta > \frac{\gamma_{\max}}{\gamma_{\min}} \cdot \theta$	$\left(0, \frac{2\delta}{\theta+\delta}\right)$
	local convergence on $D^i := \tau^i J(\rho)$	Assumption 2.5 holds at the local minimizer with $\frac{\delta^2}{2\delta+\theta} > -\min\{0, \lambda_{\min}(J(F_\phi(V^*)))\}$	$\left(0, \frac{2(\delta+\lambda^*_{\min})}{\theta+\delta}\right)$
	local convergence on $D^i := \tau^i L^\dagger$	Assumption 2.5 holds at the local minimizer with $\delta > \theta$	$\left(0, \frac{2\delta}{\theta+\delta}\right)$

derstanding of the KS minimization problem, the KS equation, and the SCF iteration. A summary of our convergence results is presented in Table 3.

Acknowledgments. The authors would like to thank Dr. Chao Yang and Prof. Aihui Zhou for discussion on the KS equation and the SCF iteration. The authors are grateful to Prof. Karl Kunisch and two anonymous referees for their detailed and valuable comments and suggestions.

REFERENCES

[1] S. L. ADLER, *Quantum theory of the dielectric constant in real solids*, Phys. Rev. (2), 126 (1962), pp. 413–420.
[2] J. F. BONNANS AND A. SHAPIRO, *Perturbation Analysis of Optimization Problems*, Springer, New York, 2000.
[3] C. LE BRIS, *Computational chemistry from the perspective of numerical analysis*, Acta Numer., 14 (2005), pp. 363–444.
[4] E. CANCÈS, *SCF algorithms for HF electronic calculations*, in Mathematical Models and Methods for Ab Initio Quantum Chemistry, Lect. Notes in Chem. 74, Springer, Berlin, 2000, pp. 17–43.
[5] E. CANCÈS, *Self-consistent field algorithms for Kohn–Sham models with fractional occupation numbers*, J. Chem. Phys., 114 (2001), pp. 10616–10622.
[6] E. CANCÈS AND C. LE BRIS, *Can we outperform the DIIS approach for electronic structure calculations?*, Int. J. Quantum Chem., 79 (2000), pp. 82–90.
[7] E. CANCÈS AND C. LE BRIS, *On the convergence of SCF algorithms for the Hartree-Fock equations*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 749–774.
[8] E. CANCÈS, M. DEFRANCESCHI, W. KUTZELNIGG, C. LE BRIS, AND Y. MADAY, *Computational quantum chemistry: A primer*, in Handbook of Numerical Analysis. Vol. X: Special Volume: Computational Chemistry, Ph. Ciarlet and C. Le Bris, eds., North-Holland, 2003, pp. 3–270.
[9] E. CANCÈS AND K. PERNAL, *Projected gradient algorithms for Hartree-Fock and density matrix functional theory calculations*, J. Chem. Phys., 128 (2008), pp. 108–134.

[10] H. Chen, X. Dai, X. Gong, L. He, and A. Zhou, *Adaptive finite element approximations for Kohn–Sham models*, Multiscale Model. Simul. 12 (2014), pp. 1828–1869.
[11] H. Chen, X. Gong, L. He, Z. Yang, and A. Zhou, *Numerical analysis of finite dimensional approximations of Kohn–Sham models*, Adv. Comput. Math., 38 (2013), pp. 225–256.
[12] X. Chen, H. Qi, and P. Tseng, *Analysis of nonsmooth symmetric-matrix-valued functions with applications to semidefinite complementarity problems*, SIAM J. Optim., 13 (2003), pp. 960–985.
[13] X. Dai, X. Gong, Z. Yang, D. Zhang, and A. Zhou, *Finite volume discretizations for eigenvalue problems with applications to electronic structure calculations*, Multiscale Model. Simul., 9 (2011), pp. 208–240.
[14] X. Dai, Z. Yang, and A. Zhou, *Symmetric finite volume schemes for eigenvalue problem in arbitrary dimensions*, Sci. China Ser. A., 51 (2008), pp. 1401–1414.
[15] C. Ding, *An Introduction to a Class of Matrix Optimization Problems*, Ph.D. thesis, National University of Singapore, Singapore, 2012.
[16] D. Drusvyatskiy and A. S. Lewis, *Tilt stability, uniform quadratic growth, and strong metric regularity of the subdifferential*, SIAM J. Optim., 23 (2013), pp. 256–267.
[17] W. Gao, C. Yang, and J. Meza, *Solving a class of nonlinear eigenvalue problems by Newton's method*, Technical report LBNL-2187 E, Lawrence Berkeley National Laboratory, Berkeley, CA, 2009.
[18] G. P. Kerker, *Efficient iteration scheme for self-consistent pseudopotential calculations*, Phys. Rev. B (3), 23 (1981), pp. 3082–3084.
[19] J. Koutecký and V. Bonacic, *On the convergence difficulties in the iterative Hartree-Fock procedure*, J. Chem. Phys., 55 (1971), pp. 2408–2413.
[20] G. Kresse and J. Furthmuller, *Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set*, Comput. Mater. Sci., 6 (1996), pp. 15–50.
[21] K. N. Kudin, G. E. Scuseria, and E. Cancès, *A black-box self-consistent field convergence algorithm: One step closer*, J. Chem. Phys., 116 (2002), pp. 8255–8261.
[22] A. Levitt, *Convergence of gradient-based algorithms for the Hartree-Fock equations*, ESAIM: Math. Model. Numer. Anal., 46 (2012), pp. 1321–1336.
[23] A. S. Lewis and H. S. Sendov, *Twice differentiable spectral functions*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 368–386.
[24] L. Lin and C. Yang, *Elliptic preconditioner for accelerating the self consistent field iteration in Kohn–Sham density functional theory*. SIAM J. Sci. Comput., 35 (2013), pp. S277–S298.
[25] X. Liu, X. Wang, Z. Wen, and Y. Yuan, *On the Convergence of the Self-Consistent Field Iteration in Kohn–Sham Density Functional Theory*, SIAM J. Matrix Anal. Appl., 35 (2014), pp. 546–558.
[26] J. Nocedal and S. Wright, *Numerical Optimization*, Springer, New York, 2006.
[27] R. Schneider, T. Rohwedder, A. Neelov, and J. Blauert, *Direct minimization for calculating invariant subspaces in density functional computations of the electronic structure*, J. Comput. Math., 27 (2009), pp. 360–393.
[28] A. Shapiro, *On Differentiability of Symmetric Matrix Valued Functions*, preprint, optimization online:2002/07/499, 2002.
[29] M. Torki, *Second-order directional derivatives of all eigenvalues of a symmetric matrix*, Nonlinear Anal., 46 (2001), pp. 1133–1150.
[30] Z. Wen, A. Milzarek, M. Ulbrich, and H. Zhang, *Adaptive regularized self-consistent field iteration with exact Hessian for electronic structure calculation*, SIAM J. Sci. Comput., 35 (2013), pp. A1299–A1324.
[31] N. Wiser, *Dielectric constant with local field effects included*, Phys. Rev. (2), 129 (1963), pp. 62–69.
[32] C. Yang, W. Gao, and J. C. Meza, *On the convergence of the self-consistent field iteration for a class of nonlinear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 30 (2009), pp. 1773–1788.
[33] C. Yang, J. C. Meza, B. Lee, and L.-W. Wang, KSSOLV—*a MATLAB toolbox for solving the Kohn–Sham equations*, ACM Trans. Math. Software, 36 (2009), 10.
[34] L.-H. Zhang and R.-C. Li, *Maximization of the sum of the trace ratio on the Stiefel manifold*, I: Theory, Science China Math., 57 (2014), pp. 2495–2508.
[35] L.-H. Zhang and R.-C. Li, *Maximization of the sum of the trace ratio on the Stiefel manifold*, II: Computation, Science China Math., 58 (2015), pp. 1549–1566.