

On the Convergence of the Self-Consistent Field Iteration in Kohn-Sham Density Functional Theory

Xin Liu^{*} Xiao Wang[†] Zaiwen Wen[‡] Yaxiang Yuan[§]

December 2, 2013

Abstract. It is well known that the self-consistent field (SCF) iteration for solving the Kohn-Sham (KS) equation often fails to converge, yet there is no clear explanation. In this paper, we investigate the SCF iteration from the perspective of minimizing the corresponding KS total energy functional. By analyzing the second-order Taylor expansion of the KS total energy functional and estimating the relationship between the Hamiltonian and the part of the Hessian which is not used in the SCF iteration, we are able to prove global convergence from an arbitrary initial point and local linear convergence from an initial point sufficiently close to the solution of the KS equation under assumptions that the gap between the occupied states and unoccupied states is sufficiently large and the second-order derivatives of the exchange correlation functional are uniformly bounded from above. Although these conditions are very stringent and are almost never satisfied in reality, our analysis is interesting in the sense that it provides a qualitative prediction of the behavior of the SCF iteration.

Key words. self-consistent field iteration, Kohn-Sham equation, Kohn-Sham total energy functional, nonlinear eigenvalue problem, global convergence, local convergence rate

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

1 Introduction

Consider the discretized Kohn-Sham (KS) equation

$$\begin{aligned} H(X)X &= X\Lambda, \\ X^T X &= I, \end{aligned} \tag{1}$$

where $X \in \mathbb{R}^{n \times k}$, the discretized Hamiltonian $H(X) \in \mathbb{R}^{n \times n}$ is a matrix function with respect to X such that $H(X)X$ is equal to the gradient of some discretized total energy functional $E(X)$ (to be defined in section 2), and $\Lambda \in \mathbb{R}^{k \times k}$ is a diagonal matrix consisting of k smallest eigenvalues of $H(X)$. The discretized KS equation is a fundamental nonlinear eigenvalue problem arising from the density functional theory (DFT) for electronic structure calculations [16, 19], in which the discretized charge density of electrons is defined as

$$\rho(X) := \text{diag}(XX^T), \tag{2}$$

^{*}State Key Laboratory of Scientific and Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, CHINA (liuxin@lsec.cc.ac.cn). Research supported in part by NSFC grants 11101409 and 11331012, and the National Center for Mathematics and Interdisciplinary Sciences, CAS.

[†]School of Mathematical Sciences, University of Chinese Academy of Sciences, CHINA (wangxiao@ucas.ac.cn). Research supported in part by Postdoc Grant 119103S175, UCAS president grant Y35101AY00, and NSFC grant 11301505.

[‡]Beijing International Center for Mathematical Research, Peking University, CHINA (wenzw@math.pku.edu.cn). Research supported in part by NSFC grants 11101274, 11322109 and 91330202.

[§]State Key Laboratory of Scientific and Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, CHINA (yyx@lsec.cc.ac.cn). Research supported in part by NSFC grant 11331012.

where $\text{diag}(A)$ denotes the vector containing the diagonal elements of the matrix A . If no confusion can arise, we omit the word “discretized” before “KS” and “charge density”, etc.

The most widely used approach for solving (1) is the self-consistent field (SCF) iteration. Starting from X^0 with $(X^0)^T X^0 = I$, the SCF iteration computes the $(i+1)$ -th iterate X^{i+1} as the solution of the linear eigenvalue problem:

$$\begin{aligned} H(X^i)X^{i+1} &= X^{i+1}\Lambda^{i+1}, \\ (X^{i+1})^T X^{i+1} &= I. \end{aligned} \tag{3}$$

When the difference between two consecutive Hamiltonians is negligible, the system is said to be self-consistent and the SCF procedure is terminated. Heuristics have been proposed to accelerate and stabilize the SCF iteration. For example, the charge mixing techniques [11, 13] replace the Hamiltonian by a new matrix constructed from a linear combination of either the potential or the charge densities computed in the previous SCF iterations and a new one obtained from certain schemes.

It is well known that the basic version of SCF iteration (3) often converges slowly or fails to converge [12] even with the help of various heuristics for decades, yet a clear explanation is not available. A convergence analysis of the SCF iteration for solving the Hartree-Fock equations according to the optimal damping algorithm (ODA) is established in [5]. The interested reader is referred to [1, 2, 3, 4, 6, 7, 14] on discussing ODA and its theoretical properties. Recently, an analysis of gradient-based algorithms for the Hartree-Fock equations is proposed in [15] using Łojasiewicz inequality. Some analysis on gradient-based algorithms can also be found in [17]. In [21], the authors prove that the sequence generated by the SCF iteration converges alternatively to two limit points which do not satisfy (1) on certain type of problems. A few numerical explanations are provided in [23] by viewing the SCF iteration as an indirect procedure of minimizing a sequence of quadratic surrogates. A condition is identified in [21] to guarantee that the SCF iteration becomes a contractive fixed point iteration under a specific form of the Hamiltonian without involving any exchange correlation term. Basically, the condition characterizes the contribution of the nonlinear component of the Hamiltonian.

In this paper, we establish some conditions on ensuring global and local convergence of the SCF iteration for general Kohn-Sham DFT from an optimization point of view. Actually, the KS equation (1) is closely related to the constrained minimization problem with orthogonality constraints

$$\begin{aligned} \min_{X \in \mathbb{R}^{n \times k}} \quad & E(X) \\ \text{s. t.} \quad & X^T X = I. \end{aligned} \tag{4}$$

The first-order optimality conditions for (4) are the same as (1) except that the diagonal matrix Λ consists of any k eigenvalues of $H(X)$ rather than the k smallest ones. Assume that the second-order derivative of the exchange correlation energy functional is uniformly bounded from above, which implies the Lipschitz continuity of the Jacobian of the functional. Inspired by the expression of the exact Hessian of $E(X)$ discovered in [9, 20], we observe that the SCF iteration discards a “complicate” term in the Hessian of the total energy functional $E(X)$. Our analysis shows that this term plays an important role in the performance of the SCF scheme (3). Briefly speaking, it converges if the gap between the k th and $(k+1)$ st eigenvalues of the Hamiltonian $H(X)$ outweighs the norm of the complicate term in the Hessian up to some constant. Although this condition is very stringent and is almost never satisfied in practice, which explains why the simplest SCF iteration often does not converge, our presented analysis is interesting theoretically in the sense that it provides a qualitative prediction of the behavior of the SCF iteration with respect to the spectral gap of the nonlinear Hamiltonian relative to the Coulomb interaction.

The rest of this paper is organized as follows. In section 2, we describe the total energy functional and its gradient

and Hessian, as well as the distance measurements between subspaces in detail. The global and local convergence of the SCF iteration are presented in section 3 and 4, respectively. Some relationship to the condition in [21] is clarified in section 5. Finally, we conclude our paper in the last section.

2 Problem Statement

2.1 The KSDFT Total Energy Functional

Consider the discretized KS total energy functional based on plane wave discretization as

$$E(X) := \frac{1}{4}\text{tr}(X^T L X) + \frac{1}{2}\text{tr}(X^T V_{ion} X) + \frac{1}{2} \sum_i \sum_l |x_i^T w_l|^2 + \frac{1}{4} \rho^T L^\dagger \rho + \frac{1}{2} e^T \epsilon_{xc}(\rho), \quad (5)$$

where $X = [x_1, \dots, x_k] \in \mathbb{R}^{n \times k}$. The first term of (5) is the so-called kinetic energy, where L is a finite dimensional representation of the Laplacian operator. The second term denotes local ionic potential energy, where the diagonal matrix V_{ion} is the ionic pseudopotentials sampled on the suitably chosen Cartesian grid. The third term defines the nonlocal ionic potential energy, where w_l represents a discretized pseudopotential reference projection function. The matrix L^\dagger corresponds to the pseudo-inverse of L and the fourth term denotes the Hartree potential energy, which is used to model the classical electrostatic average interaction between electrons. The final term denotes the exchange correlation energy, which is used to describe the nonclassical interaction between electrons. More detailed description of each terms of $E(X)$ can be found in [22, 23]. Although the function (5) is can be different if other basis functions, such as Gaussian atomic orbitals, are used for the discretization, our analysis still holds with some obvious modifications.

It can be verified that the gradient of $E(X)$ with respect to X is $\nabla E(X) = H(X)X$, where the Hamiltonian

$$H(X) := \frac{1}{2}L + V_{ion} + \sum_l w_l w_l^T + \text{Diag}(L^\dagger \rho) + \text{Diag}(\mu_{xc}(\rho)^T e), \quad (6)$$

and $\mu_{xc}(\rho) := \frac{\partial \epsilon_{xc}}{\partial \rho} \in \mathbb{R}^{n \times n}$ and $\text{Diag}(x)$ (with an uppercase letter D) denotes a diagonal matrix with x on its diagonal. Let $\mathcal{L}(\mathbb{R}^{n \times k}, \mathbb{R}^{n \times k})$ denote the space of linear operators which map $\mathbb{R}^{n \times k}$ to $\mathbb{R}^{n \times k}$. The Fréchet derivative of $\nabla E(X)$ is defined as the (unique) function $\nabla^2 E : \mathbb{R}^{n \times k} \rightarrow \mathcal{L}(\mathbb{R}^{n \times k}, \mathbb{R}^{n \times k})$ such that

$$\lim_{\|S\|_F \rightarrow 0} \frac{\|\nabla E(X+S) - \nabla E(X) - \nabla^2 E(X)(S)\|_F}{\|S\|_F} = 0.$$

The next lemma shows an explicit form of the Hessian operator [9, 20].

Lemma 2.1 (Lemma 2.1 in [20]). *Suppose that $\epsilon_{xc}(\rho(X))$ is twice differentiable with respect to $\rho(X)$. Given a direction $S \in \mathbb{R}^{n \times k}$, the Hessian-vector product of $E(X)$ is*

$$\nabla^2 E(X)[S] = H(X)S + B(X)[S], \quad (7)$$

where $J := L^\dagger + \frac{\partial^2 \epsilon_{xc}}{\partial \rho^2} e$ and

$$B(X)[S] := 2\text{Diag}(J \text{diag}(S X^T)) X. \quad (8)$$

We make the following assumptions on the total energy function.

Condition 2.2. *The second-order derivatives of the exchange correlation functional $\epsilon_{xc}(\rho)$ is uniformly bounded from above, which implies the Lipschitz continuity of its Jacobian. Without loss of generality, we assume that there exists a constant σ such that*

$$\|\text{Diag}(\mu_{xc}(\rho)^T e) - \text{Diag}(\mu_{xc}(\tilde{\rho})^T e)\|_F \leq \sigma \|\rho - \tilde{\rho}\|_2 \text{ and } \left\| \frac{\partial^2 \epsilon_{xc}}{\partial \rho^2} e \right\|_2 \leq \sigma, \quad \text{for all } \rho \in \mathbb{R}^n.$$

We next consider the second part of the Hessian operator $B(X)[S]$ defined in (8).

Lemma 2.3. *Suppose that Condition 2.2 holds. Let $X \in \mathcal{O}^{n \times k}$, $Z \in \mathcal{O}^{n \times (n-k)}$ and $S \in \mathbb{R}^{n \times k}$. Then*

$$\|B(X)[S]\|_F \leq 2\sqrt{n}(\|L^\dagger\|_2 + \sigma) \cdot \|S\|_2, \quad (9)$$

$$\|Z^T B(X)[ZZ^T S]\|_F \leq 2\sqrt{n}(\|L^\dagger\|_2 + \sigma) \cdot \|Z^T S\|_2. \quad (10)$$

Proof. We only prove the second inequality. Using $\|Z^T\|_2 \leq 1$ and $\|X\|_2 = 1$, we obtain

$$\begin{aligned} \|Z^T B(X)[ZZ^T S]\|_F &= \|2Z^T \text{Diag}(J \text{diag}(ZZ^T S X^T))X\|_F \\ &\leq 2\|Z^T\|_2 \|\text{Diag}(J \text{diag}(ZZ^T S X^T))\|_F \|X\|_2 \\ &\leq 2\|\text{Diag}(J \text{diag}(ZZ^T S X^T))\|_F = 2\|J \text{diag}(ZZ^T S X^T)\|_2 \\ &\leq 2\|J\|_2 \cdot \|\text{diag}(ZZ^T S X^T)\|_2 \leq 2\|J\|_2 \cdot \sqrt{n} \|ZZ^T S X^T\|_\infty \\ &\leq 2\sqrt{n} \|J\|_2 \cdot \|ZZ^T S X^T\|_2 \leq 2\sqrt{n} \|J\|_2 \cdot \|Z^T S\|_2, \end{aligned}$$

where the last inequality uses the fact that $\|ZM\|_2 \leq \|M\|_2$ for any matrix $M \in \mathbb{R}^{k \times k}$. This completes the proof. \square

Our analysis also relies on the gap between the k th and $(k+1)$ st eigenvalues of $H(X)$.

Condition 2.4. *Let $\lambda_1 \leq \dots \leq \lambda_k < \lambda_{k+1} \leq \dots \leq \lambda_n$ be the eigenvalues of a symmetric matrix $H \in \mathbb{R}^{n \times n}$. There exists a gap between the k th and $(k+1)$ st eigenvalues, that is, $\lambda_{k+1} - \lambda_k \geq \delta$ for some positive constant δ .*

If Condition 2.4 holds for a sequence of matrices $\{H^i\}$ ($i = 1, 2, \dots$) whose δ is uniformly bounded away from zero, $\{H^i\}$ is said to be uniformly well posed (UWP) in [1, 21].

2.2 Distance Measurements

The SCF iteration maintains orthogonality in each iteration. The feasible set

$$\mathcal{O}^{n \times k} := \{X \mid X \in \mathbb{R}^{n \times k}, X^T X = I\}$$

is often referred to as the Stiefel manifold. The solutions of the KS equation (1), the SCF iteration (3) and the minimization problem (4) are invariant with respect to orthogonal transformations. Namely, if X is a solution, all points in the set $\{XU \mid U \in \mathbb{R}^{k \times k}, U^T U = I_k\}$ are also solutions. Hence, the Euclidean distance is not suitable to measure the distance between a feasible point to a solution or a solution set of (1). Inspired by the convergence analysis in [21], we introduce two subspaces distance measurements defined in section 4.3 of [8] for further analysis, i.e., for any $X_1, X_2 \in \mathcal{O}^{n \times k}$,

$$1. \text{ Chordal 2-norm: } \mathbf{d}_{c2}(X_1, X_2) := \min_{Q_1, Q_2 \in \mathcal{O}^{k \times k}} \|X_1 Q_1 - X_2 Q_2\|_2;$$

$$2. \text{ Projection 2-norm: } \mathbf{d}_{p2}(X_1, X_2) := \|X_1 X_1^T - X_2 X_2^T\|_2.$$

Let $U\Sigma V^T$ be the singular value decomposition of $X_1^T X_2$. It holds that

$$\mathbf{d}_{\mathbf{c}2}(X_1, X_2) = \|X_1 U - X_2 V\|_2. \quad (11)$$

Since the equivalence between $\mathbf{d}_{\mathbf{c}2}$ and $\mathbf{d}_{\mathbf{p}2}$ is not discussed in [8], we next include a proof for completeness.

Lemma 2.5. *Given any $X_1, X_2 \in \mathcal{O}^{n \times k}$, the Chordal 2-norm and Projection 2-norm satisfy*

$$\mathbf{d}_{\mathbf{c}2}(X_1, X_2) \geq \mathbf{d}_{\mathbf{p}2}(X_1, X_2) \geq \frac{\sqrt{2}}{2} \mathbf{d}_{\mathbf{c}2}(X_1, X_2). \quad (12)$$

Proof. We first consider the first inequality in (12). Let us denote $\bar{X}_1 = X_1 U$ and $\bar{X}_2 = X_2 V$, where U and V are defined in (11). Then, we observe

$$\begin{aligned} 0 &\preceq (I_k - \bar{X}_1^T \bar{X}_2)(I_k - \bar{X}_2^T \bar{X}_1) = I - \bar{X}_1^T \bar{X}_2 - \bar{X}_2^T \bar{X}_1 + \bar{X}_1^T \bar{X}_2 \bar{X}_2^T \bar{X}_1 \\ &= (2I_k - \bar{X}_1^T \bar{X}_2 - \bar{X}_2^T \bar{X}_1) - (I_k - \bar{X}_1^T \bar{X}_2 \bar{X}_2^T \bar{X}_1), \end{aligned}$$

which yields

$$\sigma_{\max}(I_k - \bar{X}_1^T \bar{X}_2 \bar{X}_2^T \bar{X}_1) \leq \sigma_{\max}(2I_k - \bar{X}_1^T \bar{X}_2 - \bar{X}_2^T \bar{X}_1). \quad (13)$$

Let $Z_2 \in \mathcal{O}^{n \times (n-k)}$ be the orthogonal complement to X_2 . The left hand side of (13) satisfies

$$\begin{aligned} \sigma_{\max}(I_k - \bar{X}_1^T \bar{X}_2 \bar{X}_2^T \bar{X}_1) &= \sigma_{\max}(\bar{X}_1^T (I_k - \bar{X}_2 \bar{X}_2^T) \bar{X}_1) = \sigma_{\max}(\bar{X}_1^T Z_2 Z_2^T \bar{X}_1) \\ &= \|Z_2^T \bar{X}_1\|_2^2 = \mathbf{d}_{\mathbf{p}2}^2(\bar{X}_1, \bar{X}_2) = \mathbf{d}_{\mathbf{p}2}^2(X_1, X_2), \end{aligned} \quad (14)$$

where the last equality holds due to Theorem 2.6.1 of [10]. It follows from (11) that the right hand side of (13) satisfies

$$\sigma_{\max}(2I_k - \bar{X}_1^T \bar{X}_2 - \bar{X}_2^T \bar{X}_1) = \|\bar{X}_1 - \bar{X}_2\|_2^2 = \mathbf{d}_{\mathbf{c}2}^2(X_1, X_2), \quad (15)$$

which together with (14) proves the first part of (12).

We now prove the second inequality of (12). According to (14) and the definitions of U and V , we obtain

$$\mathbf{d}_{\mathbf{p}2}^2(X_1, X_2) = \sigma_{\max}(I_k - \bar{X}_1^T \bar{X}_2 \bar{X}_2^T \bar{X}_1) = \sigma_{\max}(I_k - \Sigma^2). \quad (16)$$

It follows from (15) that

$$\mathbf{d}_{\mathbf{c}2}^2(X_1, X_2) = \sigma_{\max}(2I_k - \bar{X}_1^T \bar{X}_2 - \bar{X}_2^T \bar{X}_1) = \sigma_{\max}(2I_k - 2\Sigma). \quad (17)$$

Since X_1 and X_2 are orthogonal matrices, each diagonal entry of the diagonal matrix Σ is in $[0, 1]$. The proof is completed by combining (16) and (17) together. \square

Theorem 4.11 in [18] and Corollary 7.2.5 in [10] are sufficient to guarantee the convergence of the invariant subspaces corresponding to the k -smallest eigenvalues.

Lemma 2.6. *Suppose that the symmetric matrix $H \in \mathbb{R}^{n \times n}$ satisfies Condition 2.4. Let $\Delta H \in \mathbb{R}^{n \times n}$ be a symmetric perturbation to H and $X, \tilde{X} \in \mathbb{R}^{n \times k}$ be the invariant subspaces associated with the k smallest eigenvalues of H and*

$H + \Delta H$, respectively. If $\|\Delta H\|_2$ is sufficiently small, it holds that

$$\mathbf{d}_{\mathbf{p}2}(X, \tilde{X}) \leq C \cdot \|\Delta H\|_2, \quad (18)$$

where C is a parameter only related to δ in Condition 2.4.

3 Global Convergence of the SCF Iteration

In this section, we prove global convergence of the SCF iteration based on the reduction of the total energy functional between two consecutive iterates. Suppose that $X \in \mathcal{O}^{n \times k}$ is an arbitrary feasible point of (4), and Y is obtained from running one SCF iteration with X as the starting point. Namely, the columns of Y are the eigenvectors associated with the k smallest eigenvalues of $H(X)$. Such a Y is not unique because the linear eigenvalue problem is invariant with respect to the orthogonal transformation. Let $U\Sigma V^T$ be the singular value decomposition of $X^T Y$, where $U, V \in \mathcal{O}^{k \times k}$. Then it follows from (11) that $\bar{Y} := YVU^T$ satisfies

$$\|X - \bar{Y}\|_2 = \mathbf{d}_{\mathbf{c}2}(X, Y). \quad (19)$$

Due to the invariance, \bar{Y} is also a solution to the linear eigenvalue problem in the SCF iteration starting from X and $E(Y) = E(\bar{Y})$. For simplicity of notation, we call \bar{Y} as the closest SCF iterate obtained from X under the Chordal 2-norm.

The second-order Taylor expansion of $E(Y)$ at X gives

$$E(Y) = E(X) + \langle \nabla E(X), Y - X \rangle + \frac{1}{2} \langle \nabla^2 E(D_t)[Y - X], Y - X \rangle,$$

where $D_t = X + t(Y - X)$ for some $t \in (0, 1)$, and the Euclidean inner product $\langle A_1, A_2 \rangle$ between any real matrices $A_1, A_2 \in \mathbb{R}^{n \times k}$ is defined as $\text{tr}(A_1^T A_2)$. Using the formulations of the gradient $\nabla E(X) = H(X)X$ and the Hessian-vector product (7), we obtain

$$\begin{aligned} E(X) - E(Y) &= -\langle \nabla E(X), Y - X \rangle - \frac{1}{2} \langle \nabla^2 E(X)[Y - X], Y - X \rangle \\ &\quad - \frac{1}{2} \langle \nabla^2 E(D_t)[Y - X], Y - X \rangle + \frac{1}{2} \langle \nabla^2 E(X)[Y - X], Y - X \rangle \\ &= \frac{1}{2} (\langle H(X)X, X \rangle - \langle H(X)Y, Y \rangle) - R_X^{(1)}(Y, D_t) - R_X^{(2)}(Y, D_t), \end{aligned} \quad (20)$$

where

$$R_X^{(1)}(Y, D_t) := \frac{1}{2} \langle (H(D_t) - H(X))(Y - X), Y - X \rangle, \quad (21)$$

$$R_X^{(2)}(Y, D_t) := \frac{1}{2} \langle B(D_t)[Y - X], Y - X \rangle. \quad (22)$$

The first term of the right hand side in (20) corresponds to a reduction of a quadratic form of the linear eigenvalue problem in the SCF iteration. Lemma 1 in [21] ensures the following reduction.

Lemma 3.1. *Suppose that Condition 2.4 holds at $H(X)$, and Y is a solution obtained from running one SCF iteration with X as the starting point. Then we have*

$$\langle H(X)X, X \rangle - \langle H(X)Y, Y \rangle \geq \delta \cdot \mathbf{d}_{\mathbf{p}2}^2(X, Y). \quad (23)$$

We next estimate $R_X^{(1)}(Y, D_t)$ and $R_X^{(2)}(Y, D_t)$ for the reduction of $E(X) - E(Y)$.

Lemma 3.2. *Suppose that Condition 2.2 holds. Let X be an orthogonal matrix with $H(X)$ satisfying Condition 2.4, and Y be a solution obtained from running one SCF iteration with X as the starting point. Then*

$$E(X) - E(Y) \geq \frac{1}{2}\delta \cdot \mathbf{d}_{\mathbf{p}2}^2(X, Y) - k\sqrt{n}(\|L^\dagger\|_2 + \sigma) \cdot (\mathbf{d}_{\mathbf{c}2}^2(X, Y) + \mathbf{d}_{\mathbf{c}2}^3(X, Y)). \quad (24)$$

Proof. Let \bar{Y} be the closest SCF iterate obtained from X under the Chordal 2-norm. Using the facts that the second term of the left hand side in (23) is invariant with respect to orthogonal transformation on Y and $\mathbf{d}_{\mathbf{p}2}(X, Y) = \mathbf{d}_{\mathbf{p}2}(X, \bar{Y})$, we obtain

$$\langle H(X)X, X \rangle - \langle H(X)\bar{Y}, \bar{Y} \rangle \geq \delta \cdot \mathbf{d}_{\mathbf{p}2}^2(X, \bar{Y}). \quad (25)$$

Simple calculations show that

$$\|XX^T - D_t D_t^T\|_2 \leq 2\|X - D_t\|_2 \leq 2\|\bar{Y} - X\|_2. \quad (26)$$

The definition of $H(X)$, Condition 2.2 and the inequality (26) give

$$\begin{aligned} & \|H(D_t) - H(X)\|_F \\ &= \|\text{Diag}(L^\dagger(\rho(X) - \rho(D_t)))\|_F + \|\text{Diag}(\mu_{xc}(\rho(X))^T e) - \text{Diag}(\mu_{xc}(\rho(D_t))^T e)\|_F \\ &\leq (\|L^\dagger\|_2 + \sigma)\|\rho(X) - \rho(D_t)\|_2 \\ &\leq \sqrt{n}(\|L^\dagger\|_2 + \sigma)\|\text{diag}(XX^T) - \text{diag}(D_t D_t^T)\|_\infty \\ &\leq \sqrt{n}(\|L^\dagger\|_2 + \sigma)\|XX^T - D_t D_t^T\|_2 \\ &\leq 2\sqrt{n}(\|L^\dagger\|_2 + \sigma)\|\bar{Y} - X\|_2, \end{aligned}$$

which further yields

$$\begin{aligned} R_X^{(1)}(\bar{Y}, D_t) &\leq \left| \frac{1}{2} \langle (H(D_t) - H(X))(\bar{Y} - X), \bar{Y} - X \rangle \right| \\ &\leq \frac{1}{2} \|H(D_t) - H(X)\|_F \|\bar{Y} - X\|_2 \|\bar{Y} - X\|_F \\ &\leq k\sqrt{n}(\|L^\dagger\|_2 + \sigma) \|\bar{Y} - X\|_2^3. \end{aligned} \quad (27)$$

It follows from (9) in Lemma 2.3 that

$$\begin{aligned} \langle B(D_t)[\bar{Y} - X], \bar{Y} - X \rangle &\leq \|B(D_t)[\bar{Y} - X]\|_F \|\bar{Y} - X\|_F \\ &\leq 2\sqrt{n}\|J\|_2 \|D_t(\bar{Y} - X)^T\|_2 \cdot k \cdot \|\bar{Y} - X\|_2 \\ &\leq 2k\sqrt{n}(\|L^\dagger\|_2 + \sigma) \|\bar{Y} - X\|_2^2, \end{aligned}$$

where the last inequality is implied by $\|D_t\|_2 = \|X + t(\bar{Y} - X)\|_2 \leq 1$. Consequently, we have

$$R_X^{(2)}(\bar{Y}, D_t) \leq \left| \frac{1}{2} \langle B(D_t)[\bar{Y} - X], \bar{Y} - X \rangle \right| \leq k\sqrt{n}(\|L^\dagger\|_2 + \sigma) \|\bar{Y} - X\|_2^2. \quad (28)$$

Substituting (25), (27) and (28) into (20), we obtain

$$E(X) - E(\bar{Y}) \geq \frac{1}{2}\delta \cdot \mathbf{d}_{\mathbf{p2}}^2(X, \bar{Y}) - k\sqrt{n}(\|L^\dagger\|_2 + \sigma)(\|X - \bar{Y}\|_2^2 + \|X - \bar{Y}\|_2^3). \quad (29)$$

Finally, the inequality (24) is proved by using (19), $\mathbf{d}_{\mathbf{p2}}(X, Y) = \mathbf{d}_{\mathbf{p2}}(X, \bar{Y})$ and $E(Y) = E(\bar{Y})$. \square

We now present our global convergence results based on the reduction of the total energy function in Lemma 3.2 and the relationship between the distance measurements in Lemma 2.5.

Theorem 3.3. *Suppose that Condition 2.2 holds. Let $\{X^i\}$ be a sequence generated by the SCF iteration such that $\{H(X^i)\}$ is uniformly well posed with a constant δ . Then $\{X^i\}$ converges to a solution to the KS equation (1), if*

$$\delta > 12k\sqrt{n}(\|L^\dagger\|_2 + \sigma). \quad (30)$$

Proof. It follows from Lemma 2.5 and Lemma 3.2 that, for any $i = 1, 2, \dots$,

$$\begin{aligned} E(X^i) - E(X^{i+1}) &\geq \left(\frac{1}{4}\delta - k\sqrt{n}(\|L^\dagger\|_2 + \sigma) \right) \mathbf{d}_{\mathbf{c2}}^2(X^i, X^{i+1}) \\ &\quad - k\sqrt{n}(\|L^\dagger\|_2 + \sigma) \mathbf{d}_{\mathbf{c2}}^3(X^i, X^{i+1}). \end{aligned} \quad (31)$$

Since X^i and X^{i+1} are both orthogonal matrices, we have

$$\mathbf{d}_{\mathbf{c2}}(X^i, X^{i+1}) \leq \|X^i\|_2 + \|X^{i+1}\|_2 = 2. \quad (32)$$

Substituting (32) into (31), we obtain

$$E(X^i) - E(X^{i+1}) \geq \left(\frac{1}{4}\delta - 3k\sqrt{n}(\|L^\dagger\|_2 + \sigma) \right) \mathbf{d}_{\mathbf{c2}}^2(X^i, X^{i+1}). \quad (33)$$

By summing (33) over all indices from 0 to i , we obtain

$$E(X^{i+1}) \leq E(X^0) - \left(\frac{1}{4}\delta - 3k\sqrt{n}(\|L^\dagger\|_2 + \sigma) \right) \sum_{j=0}^i \mathbf{d}_{\mathbf{c2}}^2(X^j, X^{j+1}). \quad (34)$$

Since $E(X^i)$ is bounded below, we have that $E(X^0) - E(X^{i+1})$ is less than some positive constant for all i . Hence, by taking limits in (34), we have

$$\lim_{i \rightarrow \infty} \mathbf{d}_{\mathbf{c2}}(X^i, X^{i+1}) = 0. \quad (35)$$

Namely, $\{X^i\}$ converges. Let

$$X^* := \lim_{i \rightarrow \infty} X^i, \quad (36)$$

and \tilde{X} be consisted of the eigenvectors associated with the k smallest eigenvalues of $H(X^*)$. It follows from Lemma 2.6 that

$$\mathbf{d}_{\mathbf{p2}}(X^{i+1}, \tilde{X}) \leq C \cdot \|H(X^i) - H(X^*)\|_2. \quad (37)$$

Taking limit on both sides and using the continuity of $H(X)$, we obtain

$$0 \leq \mathbf{d}_{\mathbf{p}2}(X^*, \tilde{X}) = \lim_{i \rightarrow \infty} \mathbf{d}_{\mathbf{p}2}(X^{i+1}, \tilde{X}) \leq \lim_{i \rightarrow \infty} C \cdot \|H(X^i) - H(X^*)\|_2 = 0. \quad (38)$$

Namely, $X^* = \tilde{X}$, which completes the proof. \square

Theorem 3.3 guarantees the convergence of the SCF iteration to a solution of the KS equation, which is more than the first-order optimality conditions for (4). In fact, when the inequality (30) holds, the reduction of the total energy (33) implies that any global minimizer of (4) is a solution of the KS equation.

4 Local Convergence of the SCF Iteration

In this section, we establish local convergence of the SCF iteration by exposing the relationship between two consecutive iterates in terms of their distances to a particular solution of (1). The results are called local analysis since it relies on the Taylor expansion in a small neighborhood of that optimal solution.

Lemma 4.1. *Suppose that Conditions 2.2 holds. Let X^* be a solution to the KS equation (1) whose $H(X^*)$ satisfies Condition 2.4, $X \in \mathcal{O}^{n \times k}$ be in a sufficiently small neighborhood of X^* , and Y be a solution obtained from running one SCF iteration with X as the starting point. Then $\mathbf{d}_{\mathbf{p}2}(X^*, Y)$ is of the same order of $\mathbf{d}_{\mathbf{p}2}(X^*, X)$, namely*

$$\mathbf{d}_{\mathbf{p}2}(X^*, Y) = O(\mathbf{d}_{\mathbf{p}2}(X^*, X)). \quad (39)$$

Proof. Using the continuity of $H(X)$, the fact that X is in a sufficiently small neighborhood of X^* and Lemma 2.6, we obtain

$$\mathbf{d}_{\mathbf{p}2}(X^*, Y) \leq C \cdot \|H(X) - H(X^*)\|_2 = O(\|X - X^*\|_2), \quad (40)$$

which proves (39). \square

Theorem 4.2. *Suppose that Conditions 2.2 holds. Let X^* be a solution to the KS equation (1) whose $H(X^*)$ satisfies Condition 2.4, X be in a sufficient small neighborhood of X^* , and Y be a solution obtained from running one SCF iteration with X as the starting point. Then*

$$\mathbf{d}_{\mathbf{p}2}(X^*, Y) \leq \frac{2\sqrt{n}(\|L^\dagger\|_2 + \sigma)}{\delta} \cdot \mathbf{d}_{\mathbf{p}2}(X^*, X) + O(\mathbf{d}_{\mathbf{p}2}^2(X^*, X)). \quad (41)$$

Proof. For convenience of exposition, we introduce $\Delta X := X^* - X$ and $\Delta Y := X^* - Y$. Recalling the fact that $\nabla E(X) = H(X)X$, we obtain the first-order Taylor expansion of $\nabla E(X^*)$ at X as follows,

$$\begin{aligned} H(X^*)X^* = \nabla E(X^*) &= \nabla E(X) + \nabla^2 E(X)[\Delta X] + O(\|\Delta X\|_2^2) \\ &= H(X)X + H(X)\Delta X + B(X)[\Delta X] + O(\|\Delta X\|_2^2) \\ &= H(X)Y + H(X)\Delta Y + B(X)[\Delta X] + O(\|\Delta X\|_2^2). \end{aligned} \quad (42)$$

Using Lemma 4.1 and substituting X^* by $Y + \Delta Y$, we have

$$\begin{aligned} X^*(X^*)^T H(X^*)X^* &= (Y + \Delta Y)(Y + \Delta Y)^T (H(X)Y + H(X)\Delta Y + B(X)[\Delta X] + O(\|\Delta X\|_2^2)) \\ &= YY^T H(X)Y + Y\Delta Y^T H(X)Y + \Delta Y Y^T H(X)Y \\ &\quad + YY^T H(X)\Delta Y + YY^T B(X)[\Delta X] + O(\|\Delta X\|_2^2). \end{aligned} \quad (43)$$

By using the fact that X^* is a global solution of (1) and Y is an SCF iterate obtained from X , we have

$$H(X^*)X^* = X^*(X^*)^T H(X^*)X^*, \quad (44)$$

$$H(X)Y = YY^T H(X)Y. \quad (45)$$

It follows from the relations (42)-(45) that

$$\begin{aligned} &H(X)\Delta Y - (Y\Delta Y^T H(X)Y + \Delta Y Y^T H(X)Y + YY^T H(X)\Delta Y) \\ &= -(I - YY^T)B(X)[\Delta X] + O(\|\Delta X\|_2^2). \end{aligned} \quad (46)$$

Consequently, the above relation and Lemma 4.1 imply that

$$\begin{aligned} &H(X^*)\Delta Y - (X^*\Delta Y^T H(X)Y + \Delta Y(X^*)^T H(X^*)X^* + X^*Y^T H(X)\Delta Y) \\ &= -(I - X^*(X^*)^T)B(X)[\Delta X] + O(\|\Delta X\|_2^2). \end{aligned} \quad (47)$$

Let Z^* be the orthogonal complement to X^* . Multiplying both sides of (47) with $(Z^*)^T$ yields:

$$\begin{aligned} &(Z^*)^T H(X^*)\Delta Y - (Z^*)^T (X^*\Delta Y^T H(X)Y + \Delta Y(X^*)^T H(X^*)X^* + X^*Y^T H(X)\Delta Y) \\ &= -(Z^*)^T B(X)[\Delta X] + (Z^*)^T X^*(X^*)^T B(X)[\Delta X] + O(\|\Delta X\|_2^2), \end{aligned} \quad (48)$$

which can be rewritten as

$$(Z^*)^T H(X^*)\Delta Y - (Z^*)^T \Delta Y(X^*)^T H(X^*)X^* = -(Z^*)^T B(X)[\Delta X] + O(\|\Delta X\|_2^2). \quad (49)$$

Let Λ_k and Λ_{n-k} be the diagonal matrices consisting of the k smallest and $n - k$ largest eigenvalues of $H(X^*)$, respectively. It follows from (44) and the definition of Z^* that

$$\Lambda_{n-k}(Z^*)^T \Delta Y - (Z^*)^T \Delta Y \Lambda_k = -(Z^*)^T B(X)[(Z^*(Z^*)^T + X^*(X^*)^T)\Delta X] + O(\|\Delta X\|_2^2). \quad (50)$$

By using the orthogonality of X , we have $(X^* - \Delta X)^T(X^* - \Delta X) = X^T X = I$, which further gives,

$$(X^*)^T \Delta X = O(\|\Delta X\|^2). \quad (51)$$

It follows from (51) that

$$\Lambda_{n-k}(Z^*)^T \Delta Y - (Z^*)^T \Delta Y \Lambda_k = -(Z^*)^T B(X)[Z^*(Z^*)^T \Delta X] + O(\|\Delta X\|_2^2). \quad (52)$$

Taking Frobenius-norm on both sides of (52), we have

$$\|\Lambda_{n-k}(Z^*)^T \Delta Y\|_F - \|(Z^*)^T \Delta Y \Lambda_k\|_F \leq \|(Z^*)^T B(X)[Z^*(Z^*)^T \Delta X]\|_F + O(\|\Delta X\|_2^2). \quad (53)$$

Condition 2.4 implies

$$\|\Lambda_{n-k}(Z^*)^T \Delta Y\|_F - \|(Z^*)^T \Delta Y \Lambda_k\|_F \geq \delta \|(Z^*)^T \Delta Y\|_F. \quad (54)$$

By using Lemma 2.3 and substituting (54) into (53), we obtain

$$\delta \|(Z^*)^T \Delta Y\|_F \leq 2\sqrt{n}\|J\|_2 \cdot \|(Z^*)^T \Delta X\|_2 + O(\|\Delta X\|_2^2). \quad (55)$$

It is clear that $\mathbf{d}_{\mathbf{p}2}(X^*, Y) = \|(Z^*)^T \Delta Y\|_2 \leq \|(Z^*)^T \Delta Y\|_F$ and $\mathbf{d}_{\mathbf{p}2}(X^*, X) = \|(Z^*)^T \Delta X\|_2$. Recalling (51) and the definition of Z^* , we obtain

$$\|\Delta X\|_2 \geq \|(Z^*)^T \Delta X\|_2 \geq \|\Delta X\|_2 - \|(X^*)^T \Delta X\|_2 = \|\Delta X\|_2 - O(\|\Delta X\|_2^2). \quad (56)$$

Namely, $O(\|\Delta X\|_2) = O(\mathbf{d}_{\mathbf{p}2}(X^*, X))$ holds, which completes the proof. \square

Hence, when $2\sqrt{n}(\|L^\dagger\|_2 + \sigma) < \delta$ holds, Theorem 4.2 implies that the SCF iteration converges linearly to the solution X^* of the KS equation once the sequence locates in a sufficiently small neighborhood of X^* .

5 Comparison with the Results of Yang et al. in [21]

In this section, we explain the difference between our convergence results and these of Yang et al. [21] on a special form of the total energy functional as

$$E(X) := \frac{1}{2}\text{tr}(X^T L X) + \frac{\alpha}{4}\rho(X)^T L^{-1} \rho(X),$$

whose Hamiltonian is

$$H(X) := L + \alpha \text{Diag}(L^{-1} \rho(X)).$$

Since there is no exchange correlation energy functional in this case, the constant $\sigma = 0$ in Condition 2.2.

Theorem 3.3 provides global convergence from any initial point if

$$\alpha < \alpha_G := \frac{\delta}{12k\sqrt{n}\|L^{-1}\|_2}. \quad (57)$$

According to Theorem 4.2, the SCF iteration converges linearly to the optimal solution from an initial point located in a neighborhood of that solution, if α satisfies

$$\alpha < \alpha_L := \frac{\delta}{2\sqrt{n}\|L^{-1}\|_2}. \quad (58)$$

On the other hand, Yang et al. [21] proves convergence of a variant of the SCF iteration whose the density function is computed by

$$\rho = \text{diag}(f_\mu(H)).$$

Here $f_\mu(t) := \frac{1}{1+e^{\beta(t-\mu)}}$ and $f_\mu(H) := V\text{Diag}(f_\mu(\lambda_1), \dots, f_\mu(\lambda_n))V^T$, where $H = V\text{Diag}(\lambda_1, \dots, \lambda_n)V^T$ is the eigenvalue decomposition of H . They provide global linear convergence if

$$\alpha < \alpha_F := \frac{2}{n^4\beta\|L^{-1}\|_1}, \quad (59)$$

where β and μ satisfy

$$\text{trace}(f_\mu(H)) = k.$$

For a given constant $\gamma \ll 1$, the smoothing can be achieved by requiring

$$\begin{cases} \frac{1}{1+e^{\beta(\lambda_k-\mu)}} \geq 1-\gamma, \\ \frac{1}{1+e^{\beta(\lambda_{k+1}-\mu)}} \leq \gamma, \end{cases}$$

which is equivalent to

$$\beta \geq \max \left\{ \frac{\ln \frac{1-\gamma}{\gamma}}{\mu - \lambda_k}, \frac{\ln \frac{1-\gamma}{\gamma}}{\lambda_{k+1} - \mu} \right\}.$$

Notice that

$$\min_{\mu} \max \left\{ \frac{\ln \frac{1-\gamma}{\gamma}}{\mu - \lambda_k}, \frac{\ln \frac{1-\gamma}{\gamma}}{\lambda_{k+1} - \mu} \right\} = \frac{2}{\delta} \cdot \ln \frac{1-\gamma}{\gamma},$$

whose minimum is achieved at $\mu = \frac{\lambda_k + \lambda_{k+1}}{2}$. Therefore, we obtain $\beta \geq \frac{2}{\delta} \cdot \ln \frac{1-\gamma}{\gamma}$. Namely,

$$\alpha_F < \frac{\delta}{\ln \frac{1-\gamma}{\gamma} \cdot n^4 \|L^{-1}\|_1}. \quad (60)$$

We notice that $k\sqrt{n} < n^{1.5} < n^4$ and $k\sqrt{n} \ll n^4$ when n is sufficiently large. Moreover, $\ln \frac{1-\gamma}{\gamma} > 12$ if $\gamma < 6.1442 \times 10^{-6}$, whereas $\ln \frac{1-\gamma}{\gamma} \cdot n^4 > 12k\sqrt{n}$, when $\gamma < 0.1070$ and $n \geq 2$. By comparing (60) to (57), we can obtain that $\alpha_F < \alpha_G$ under a reasonable value of γ . Furthermore, $\alpha_F \ll \alpha_G$ holds when n is sufficiently large. Hence, we can conclude that our condition is no more restricted than the one in [21].

6 Conclusion

We study the convergence issues of the well-known self-consistent field (SCF) iteration for solving the Kohn-Sham equation in density functional theory. Our analysis is based on the second-order Taylor expansion of the total energy functional. We show that a ‘‘complicate’’ part of the Hessian plays an important role in ensuring the convergence of the SCF iteration. Both global and local convergence can be guaranteed if the gap between the k th and $(k+1)$ th eigenvalues of the Hamiltonian $H(X)$ outweighs the norm of the complicate term in the Hessian up to some constant and if the second-order derivatives of the exchange correlation energy is uniformly bounded from above.

Although our conditions are restrictive for the convergence of the SCF iteration and they are almost never satisfied in reality, they still provide us some insights on the performance of the algorithm. Recently, numerical evidences show that the exact Hessian can speed up the convergence of the SCF iteration in the trust-region framework [20]. Our analysis has not covered the acceleration scheme using charge mixing since it is a fixed-point algorithm in terms of the charge density rather than minimizing the total energy functional.

Acknowledgements Z. Wen would like to thank Humboldt Foundation for the generous support, Prof. Michael Ulbrich for hosting his visit at Technische Universität München. The authors would like to thank Dr. Chao Yang

for discussion on the Kohn-Sham equation and are grateful to two anonymous referees for their detailed and valuable comments and suggestions.

References

- [1] C. L. BRIS, *Computational chemistry from the perspective of numerical analysis*, Acta Numer., 14 (2005), pp. 363–444.
- [2] E. CANCÈS, *Scf algorithms for hartree-fock electronic calculations*, Lecture Notes in Chemistry, 74 (2000), pp. 17–43.
- [3] ———, *Self-consistent field algorithms for kohnsham models with fractional occupation numbers*, Journal of Chemical Physics, 114(24) (2001), p. 1061610622.
- [4] E. CANCÈS AND C. L. BRIS, *Can we outperform the diis approach for electronic structure calculations?*, International Journal of Quantum Chemistry, 79(2) (2000), pp. 82–90.
- [5] E. CANCÈS AND C. L. BRIS, *On the convergence of SCF algorithms for the Hartree-Fock equations*, Math. Model. Numer. Anal., 34 (2000), pp. 749–774.
- [6] E. CANCÈS, M. DEFRANCESCHI, W. KUTZELNIGG, C. L. BRIS, AND Y. MADAY, *Handbook of numerical analysis. Volume X: special volume: computational chemistry*, North-Holland, 2003, ch. Computational quantum chemistry: a primer, pp. 3–270.
- [7] E. CANCÈS AND K. PERNAL, *Projected gradient algorithms for hartree-fock and density matrix functional theory calculations*, Journal of Chemical Physics, 128(13) (2008), pp. 108–134.
- [8] A. EDELMAN, T. ARIAS, AND S. SMITH, *The geometry of algorithms with orthogonality constraints*, SIAM J. Matrix Analysis Applications, 20(2) (1998), pp. 303–353.
- [9] W. GAO, C. YANG, AND J. MEZA, *Solving a class of nonlinear eigenvalue problems by Newton’s method*, tech. rep., Lawrence Berkeley National Laboratory, 2009.
- [10] G. GOLUB AND C. V. LOAN, *Matrix Computaion*, The Johns and Hopkins University Press, 1996.
- [11] G. P. KERKER, *Efficient iteration scheme for self-consistent pseudopotential calculations*, Phys. Rev. B, 23 (1981), pp. 3082–3084.
- [12] J. KOUTECKÝ AND V. BONACIC, *On the convergence difficulties in the iterative Hartree-Fock procedure*, J. Chem. Phys., 55 (1971), pp. 2408–2413.
- [13] G. KRESSE AND J. FURTHMULLER, *Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set*, Computational Materials Science, 6 (1996), pp. 15–50.
- [14] K. N. KUDIN, G. E. SCUSERIA, AND E. CANCÈS, *A black-box self-consistent field convergence algorithm: One step closer*, Journal of Chemical Physics, 116(19) (2002), pp. 8255–8261.
- [15] A. LEVITT, *Convergence of gradient-based algorithms for the hartree-fock equations*, ESAIM: Mathematical Modelling and Numerical Analysis, 46(6) (2012), pp. 1321–1336.

- [16] R. M. MARTIN, *Electronic Structure: Basic Theory and Practical Methods*, Cambridge University Press, 2004.
- [17] R. SCHNEIDER, T. ROHWEDDER, A. NEELOV, JOHANNES, AND BLAUERT, *Direct minimization for calculating invariant subspaces in density functional computations of the electronic structure*, Journal of Computational Mathematics, 27(2/3) (2009), pp. 360–393.
- [18] G. W. STEWART, *Error bounds for approximation invariant subspace of closed linear operators*, SIAM Review, 15 (1973), pp. 27–64.
- [19] A. SZABO AND N. S. OSTLUND, *Modern Quantum Chemistry: An Introduction to Advanced Electronic Structure Theory*, Dover, New York, 1996.
- [20] Z. WEN, A. MILZAREK, M. ULBRICH, AND H. ZHANG, *Adaptive regularized self-consistent field iteration with exact hessian for electronic structure calculation*, SIAM Journal on Scientific Computing, 35(3) (2013), pp. A1299–A1324.
- [21] C. YANG, W. GAO, AND J. MEZA, *On the convergence of the self-consistent field iteration for a class of nonlinear eigenvalue problems*, SIAM J. Matrix Analysis Applications, 30(4) (2009), pp. 1773–1788.
- [22] C. YANG, J. C. MEZA, B. LEE, AND L.-W. WANG, *KSSOLV—a MATLAB toolbox for solving the Kohn-Sham equations*, ACM Trans. Math. Softw., 36 (2009), pp. 1–35.
- [23] C. YANG, J. C. MEZA, AND L. WANG, *A trust region direct constrained minimization algorithm for the Kohn-Sham equation*, SIAM Journal of Scientific Computing, 29 (2007), pp. 1854–1875.