ADAPTIVE REGULARIZED SELF-CONSISTENT FIELD ITERATION WITH EXACT HESSIAN FOR ELECTRONIC STRUCTURE CALCULATION

ZAIWEN WEN[†], ANDRE MILZAREK[‡], MICHAEL ULBRICH[‡], AND HONGCHAO ZHANG[§]

Abstract. The self-consistent field (SCF) iteration has been used ubiquitously for solving the Kohn-Sham (KS) equation or the minimization of the KS total energy functional with respect to orthogonality constraints in electronic structure calculations. Although SCF with heuristics such as charge mixing often works remarkably well on many problems, it is well known that its convergence can be unpredictable and there is no general theoretical analysis on their performance. We regularize the SCF iteration and establish rigorous global convergence to the first-order optimality conditions. The Hessian of the total energy functional is further exploited. By adding the part of the Hessian which is not considered in SCF, our methods can always achieve a high accurate solution on problems for which SCF fails and even exhibit quadratic or superlinear convergence on most test problems in the KSSOLV toolbox under the Matlab environment.

Key words. Density functional theory, Kohn-Sham total energy minimization, Orthogonality constraints, Regularized SCF, Trust region methods

AMS subject classifications. 15A18, 65K10, 65F15, 90C26, 90C30

1. Introduction. Electronic structure calculations have became a fundamental tool in many fields, including quantum physics, quantum chemistry and material sciences. With the advance of the Hartree-Fock and Kohn-Sham density functional theory (DFT) [9, 12], molecules with more and more complex electronic structures and larger parts of potential surfaces can be simulated. In this theory, a central but very challenging task is the minimization of the electronic total energy functional with respect to electron wave functions, which are required to be orthogonal to each other for the physical interpretation of quantum states.

The most widely used approach is the so called self-consistent field (SCF) iteration, which computes a few smallest eigenvalues and their corresponding eigenvectors of the Hamiltonian matrices iteratively. It is well known that the basic form of SCF often converges slowly or fails to converge [14]. Heuristics have been proposed to accelerate and stabilize the SCF iteration. For example, the charge mixing techniques [10, 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. Although these heuristics can work remarkably well on many problems, they can still fail on many other instances and there is no general theoretical analysis on their performance [24].

During decades of research, there have been a few optimization approaches [2, 3, 16, 17, 21, 22] on minimizing the total energy functional directly. Due to the difficulties related to the orthogonality constraints, these methods are usually much slower than SCF. A direct constrained minimization (DCM) algorithm is designed in [26] where the new search direction is built from a subspace spanned by the current approximation to the optimal wavefunction, the preconditioned gradient and the previous search direction. Wen and Yin proposed a projected gradient-type method in [23] for minimizing a general function with orthogonality constraints with guaranteed convergence. It application to DFT shows that the method can often reach

[†]Department of Mathematics, MOE-LSC and Institute of Natural Sciences, Shanghai Jiaotong University, China. (zw2109@sjtu.edu.cn). Research supported in part by NSFC grant 11101274, Shanghai Pujiang Program 12PJ1404800, and Humboldt Research Fellowship for Experienced Researchers.

[‡]Chair of Mathematical Optimization, Department of Mathematics, Technische Universität München, Boltzmannstr. 3, 85747 Garching b. München, Germany. (milzarek@ma.tum.de, mulbrich@ma.tum.de).

[§]Department of Mathematics and Center for Computational & Technology at Louisiana State University. (hozhang@math.lsu.edu). Research supported in part by NSF DMS-1016204.

a low accurate solution quickly but may take a lot of iterations to achieve a high accurate solution since only the gradient information is used. Trust region methods [15, 19] have also been applied to solve DFT in [5, 6, 18, 20, 27]. These approaches substitute the linear eigenvalue problem in SCF by the so called trust-region subproblems, in which the objective function are local quadratic approximations to the total energy functional. Monotonic reduction of the total energy can be achieved by imposing a suitable update of the trust region radius. However, these trust region methods compared with SCF can still be extremely slow for large systems.

In the present paper, we aim to regularize the SCF iteration by using the trust region framework and accelerate the convergence by further exploiting the Hessian of the total energy functional. Basically, the linear eigenvalue problem in SCF can be interpreted as the minimization of a quadratic function with respect to an orthogonality matrix. Our methods add a regularization term to the approximated quadratic model of the objective function. By adaptively controlling the regularization parameter, global and fast local convergence can be established rigorously similar to the technique developed in [4, 8]. Theoretically, the exact or asymptotic accurate full Hessian would be necessary for a fast local convergence. Inspired by the expression of the exact Hessian of the total energy discovered in [7], we can observe that the quadratic approximation of SCF discards a "complicate" yet important term in the Hessian. Although the computation of this extra Hessian term itself is not practical, the corresponding Hessian-vector products can be evaluated efficiently. After adding the extra Hessian term in our regularized SCF model, the methods have in general better performance than the SCF method. They exhibit quadratic or superlinear convergence on most testing problems in the software package KSSOLV [25] and can always attain a high accuracy even on problems for which SCF fails.

This paper is organized as follows. In Section 2, we review the background on Kohn-Sham total energy minimization, present the gradient and Hessian of the total energy functional and then explain our motivation from algorithms for solving the nonlinear least-squares problems. Our regularized SCF methods are proposed in Section 3 and their convergence analysis is established in Section 4. Finally, we demonstrate the robustness and efficiency of our algorithms based on KSSOLV in Section 5.

1.1. Notation. Given $X \in \mathbb{C}^{m\times n}$, the operators $\bar{X}$, X^*, $\Re(X)$ and $\Im(X)$ denote the complex conjugate, the complex conjugate transpose, the real and imaginary parts of X, respectively. The trace of X, i.e., the sum of the diagonal elements of $X \in \mathbb{C}^{n\times n}$, is denoted by $\mathrm{tr}(X)$. The Euclidean inner product between two matrices $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{m\times n}$ is defined as $\langle A, B\rangle := \sum_{jk} A^*_{j,k}B_{j,k} = \mathrm{tr}(A^*B)$. The operations $A \otimes B$ and $A \odot B$ are the Kronecker and Hadamard products, respectively. For simplicity of notation, we write $\Re\langle A, B\rangle$ as $\langle A, B\rangle$ if the value of the latter is real. The Frobenius norm of $A \in \mathbb{C}^{m\times n}$ is defined as $\|A\|_F := \sqrt{\sum_{i,j} A^*_{i,j}A_{i,j}}$. For a given $d \in \mathbb{C}^n$, the operator $\mathrm{diag}(d)$ returns a square matrix in $\mathbb{C}^{n\times n}$ with the elements of d on the main diagonal, while $\mathrm{diag}(X)$ gives a vector in $\mathbb{C}^n$ consisting of the main diagonal of X. For a given X, the vectorization $\mathrm{vec}(X)$ is a column vector obtain by stacking the columns of the matrix X on top of one another, and the operation $\mathrm{mat}(X)$ is defined as $\mathrm{mat}(\mathrm{vec}(X)) = X$. The notation e refers to the column vector of all ones.

Suppose that $\mathcal{F}(X) : \mathbb{C}^{n\times p} \to \mathbb{R}$ is a nonholomorphic function. Then $\mathcal{F}(X)$ can also be written as $\mathcal{F}(X, \bar{X})$. If no confusion can arise, we refer $\mathcal{F}(X)$ to be independent of $\bar{X}$ and $\mathcal{F}(X, \bar{X})$ to mean a function of both X and $\bar{X}$. The R-derivative and conjugate R-derivate of $\mathcal{F}$ are taken to be standard complex partial derivatives with respect to X and $\bar{X}$, i.e.,

$$\frac{\partial \mathcal{F}(X, \bar{X})}{\partial X}\Big|_{\bar{X}=const} \text{ and } \frac{\partial \mathcal{F}(X, \bar{X})}{\partial \bar{X}}\Big|_{X=const}.$$

The differential rule is defined as

$$d\mathcal{F} = \frac{\partial \mathcal{F}(X, \bar{X})}{\partial X} dX + \frac{\partial \mathcal{F}(X, \bar{X})}{\partial \bar{X}} d\bar{X}.$$

The second-order Taylor expansion in X can be expressed as

$$\mathcal{F}(X + \Delta X) = \mathcal{F}(X) + 2\Re \left\langle \frac{\partial \mathcal{F}}{\partial \bar{X}}, \Delta X \right\rangle + \Re \left\langle \mathcal{F}_{XX} \Delta X + \mathcal{F}_{X\bar{X}} \Delta \bar{X}, \Delta X \right\rangle + h.o.t, \tag{1.1}$$

where

$$\mathcal{F}_{XX} = \frac{\partial}{\partial X} \left(\frac{\partial \mathcal{F}}{\partial X} \right)^*, \mathcal{F}_{\bar{X}X} = \frac{\partial}{\partial \bar{X}} \left(\frac{\partial \mathcal{F}}{\partial X} \right)^*, \mathcal{F}_{X\bar{X}} = \frac{\partial}{\partial X} \left(\frac{\partial \mathcal{F}}{\partial \bar{X}} \right)^* \text{ and } \mathcal{F}_{\bar{X}\bar{X}} = \frac{\partial}{\partial \bar{X}} \left(\frac{\partial \mathcal{F}}{\partial \bar{X}} \right)^*,$$

and $h.o.t.$ denotes the higer order terms. From the above definitions, the gradient of $\mathcal{F}$ is defined as

$$(\nabla \mathcal{F})_{ij} = 2 \frac{\partial \mathcal{F}}{\partial \bar{X}_{ij}}.$$

and the Hessian-vector product is defined as

$$(\nabla^2 \mathcal{F}) \Delta X = 2 \left(\mathcal{F}_{XX} \Delta X + \mathcal{F}_{X\bar{X}} \Delta \bar{X} \right).$$

Hence, the expansion (1.1) can be rewritten as

$$\mathcal{F}(X + \Delta X) = \mathcal{F}(X) + \Re \left\langle \nabla \mathcal{F}, \Delta X \right\rangle + \frac{1}{2} \Re \left\langle (\nabla^2 \mathcal{F}) \Delta X, \Delta X \right\rangle + h.o.t. \tag{1.2}$$

2. Background and Motivation.

2.1. Kohn-Sham Total Energy Minimization. In this section, we introduce minimizing the discretized Kohn-Sham (KS) total energy functional with respect to electron wave functions. Using a suitable discretization scheme, the electron wave functions of p occupied states can be approximated by a matrix

$$X = [x_1, \cdots, x_p] \in \mathbb{C}^{n \times p},$$

where n is the spatial degrees of freedom. Due to physical constraints, the wave functions X must be orthogonal to each other, namely:

$$X^* X = I,$$

where I is the identity matrix. The charge density associated with the occupied states can be expressed as

$$\rho(X) := \text{diag}(X X^*).$$

The finite-dimensional approximation to the continuous KS total energy functional is defined as

$$E(X) := \frac{1}{4} \text{tr}(X^* L X) + \frac{1}{2} \text{tr}(X^* V_{ion} X) + \frac{1}{2} \sum_i \sum_l |x_i^* w_l|^2 + \frac{1}{4} \rho^\top L^\dagger \rho + \frac{1}{2} e^\top \epsilon_{xc}(\rho). \tag{2.1}$$

Here, the first term of (2.1) 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. An example of $\epsilon_{xc}(\rho)$ is given in section 5. More detailed description of each terms of $E(X)$ can be found in [27, 25].

The discretized KS total energy minimization problem is

$$\begin{aligned} \min \quad & E(X) \\ \text{s.t.} \quad & X^*X = I. \end{aligned} \tag{2.2}$$

It can be verified that the gradient of $E(X)$ with respect to X is

$$G(X) := \nabla E(X) = H(X)X,$$

where the Hamiltonian

$$H(X) := \frac{1}{2}L + V_{ion} + \sum_l w_l w_l^* + \mathrm{diag}((\Re L^\dagger)\rho) + \mathrm{diag}(\mu_{xc}^* e), \tag{2.3}$$

and $\mu_{xc} = \frac{\partial \epsilon_{xc}}{\partial \rho} \in \mathbb{R}^{n\times n}$. Hence, the first-order optimality conditions of (2.2) are

$$\begin{aligned} H(X)X &= X\Lambda, \\ X^*X &= I, \end{aligned} \tag{2.4}$$

where Λ is the Lagrangian multiplier that is a symmetric matrix.

Equations (2.4) has been dubbed as the *Kohn-Sham equation*, which is a type of nonlinear eigenvalue problem since $H(X)$ is a function of the vector ρ or X. The most widely used technique for solving (2.2) or (2.4) is the self-consistent field (SCF) iteration. Starting from X_0 with $X_0^*X_0 = I$, SCF computes the $(k+1)$-th iterate X_{k+1} as the solution of the linear eigenvalue problem:

$$\begin{aligned} H(X_k)X_{k+1} &= X_{k+1}\Lambda_{k+1}, \\ X_{k+1}^*X_{k+1} &= I. \end{aligned} \tag{2.5}$$

The convergence of SCF can often be sped up by the so called charge mixing. The only difference is that the coefficient matrix $H(X_k)$ of the linear eigenvalue (2.5) is replaced by another matrix $\widehat{H}$, which is constructed from a linear combination of the previously computed potential or charge densities and the one obtained from certain schemes at current iteration. Frequently used schemes include Pulay (or DIIS), Broyden and Anderson mixing. Although SCF with charge mixing often works well on many problems, it is well known that there is no theoretical guarantee on its convergence and it can converge very slowly or fail on certain problems. On the other hand, the main computational bottleneck of SCF is solving a sequence of linear eigenvalue problems.

2.2. Our Motivation. We first present the Hessian of the total energy functional and then describe the connection between SCF and its optimization counterpart.

LEMMA 2.1. *Suppose that $\epsilon_{xc}(\rho(X))$ is twice differentiable with respect to $\rho(X)$. Given a direction $Z \in \mathbb{C}^{n\times p}$, the Hessian-vector of $E(X)$ is*

$$\nabla^2(E(X))[Z] = H(X)Z + \operatorname{diag}\left(J\left((\bar{X}\odot Z + X\odot \bar{Z})e\right)\right)X, \tag{2.6}$$

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

Proof. It can be verified that

$$\begin{aligned}(\nabla_{\mathrm{vec}(X)}\rho)\mathrm{vec}(Z) &= \frac{\partial \rho}{\partial \mathrm{vec}(X)}\mathrm{vec}(Z) + \frac{\partial \rho}{\partial \mathrm{vec}(\bar{X})}\mathrm{vec}(\bar{Z}) \\ &= (e^\top \otimes I_n)\left(\mathrm{diag}(\mathrm{vec}(\bar{X}))\mathrm{vec}(Z) + \mathrm{diag}(\mathrm{vec}(X))\mathrm{vec}(\bar{Z})\right).\end{aligned}$$

Since the gradient of $E(X)$ is $H(X)X$, it is suffice to compute the gradient of $\mathrm{vec}(H(X)X)$. Using similar proofs as in section 3.1 in [7], we obtain

$$\begin{aligned}&(\nabla \mathrm{vec}(H(X)X))\,\mathrm{vec}(Z) \\ &= I_p \otimes H(X)\mathrm{vec}(Z) + \mathrm{diag}(\mathrm{vec}(X))\nabla\left(e\otimes(L^\dagger\rho + \mu_{xc})\right)\mathrm{vec}(Z) \\ &= \mathrm{vec}(H(X)Z) + \mathrm{diag}(\mathrm{vec}(X))\frac{\partial\left(e\otimes(L^\dagger\rho+\mu_{xc})\right)}{\partial\rho}\frac{\partial\rho}{\partial\mathrm{vec}(X)}\mathrm{vec}(Z) \\ &= \mathrm{vec}(H(X)Z) + \mathrm{diag}(\mathrm{vec}(X))\,(e\otimes J)\,(e^\top\otimes I_n)\left(\mathrm{diag}(\mathrm{vec}(\bar{X}))\mathrm{vec}(Z) + \mathrm{diag}(\mathrm{vec}(X))\mathrm{vec}(\bar{Z})\right) \\ &= \mathrm{vec}(H(X)Z) + 2\,\mathrm{diag}(\mathrm{vec}(X))\left(ee^\top\otimes J\right)\left(\mathrm{diag}(\mathrm{vec}(\bar{X}))\mathrm{vec}(Z) + \mathrm{diag}(\mathrm{vec}(X))\mathrm{vec}(\bar{Z})\right) \\ &= \mathrm{vec}(H(X)Z) + \mathrm{vec}\left(X\odot\left(J(\bar{X}\odot Z + X\odot\bar{Z})ee^\top\right)\right)\end{aligned}$$

which gives (2.6). ∎

Note that the KS equation (2.5) corresponds to the first-order optimality conditions of the subproblem

$$\begin{aligned}\min \quad & q(X) := \frac{1}{2}\langle H(X_k)X, X\rangle, \\ \text{s.t.} \quad & X^*X = I.\end{aligned} \tag{2.7}$$

On the other hand, a direct calculation reveals:

$$\Re\langle H(X_k)X_k, X - X_k\rangle + \frac{1}{2}\Re\langle H(X_k)(X-X_k), X - X_k\rangle = \frac{1}{2}\langle H(X_k)X, X\rangle + \text{constant}. \tag{2.8}$$

Comparing (2.8) with the Hessian operator (2.6), the quadratic surrogate function $q(X)$ is a second-order Taylor expansion of $E(X)$ without involving the second term in the right hand side of (2.6).

We are motivated to improve SCF by the Gauss-Newton, Levenberg-Marquardt and Quasi-Newton methods for solving the nonlinear least squares problems [15, 19]:

$$\min_{x\in\mathbb{R}^n} f(x) = \frac{1}{2}\sum_{j=1}^{m} r_j^2(x) = \|r(x)\|_2^2, \tag{2.9}$$

where $r(x) = [r_1(x), \ldots, r_m(x)]^\top$ and $r_j(x) : \mathbb{R}^n \to \mathbb{R}$. The gradient and Hessian of $f(x)$ are

$$\nabla f(x) = \sum_{j=1}^{m} r_j(x) \nabla r_j(x) = J(x)^\top r(x),$$

$$\nabla^2 f(x) = J(x)^\top J(x) + \sum_{j=1}^{m} r_j(x) \nabla^2 r_j(x),$$

where the Jacobian matrix $J(x) = [\nabla r_1(x)^\top, \ldots, \nabla r_m(x)^\top]$. These algorithms update the k-th iteration x_k as follows.

1. Newton's method: $x_{k+1} = x_k + \alpha_k d_k$, where α_k is a stepsize and d_k is the solution of

$$\nabla^2 f(x_k) d_k = -\nabla f(x_k).$$

2. Gauss-Newton method, i.e., an approximate Newton method without the term $\sum_{j=1}^{m} r_j(x) \nabla^2 r_j(x)$: $x_{k+1} = x_k + \alpha_k d_k$, where α_k is a stepsize and d_k is the solution of

$$J_k^\top J_k d_k = -\nabla f(x_k).$$

3. Levenberg-Marquardt Method: $x_{k+1} = x_k + d_k$, where d_k is the solution of

$$\min_{d \in \mathbb{R}^n} \; d^\top \nabla f(x_k) + \frac{1}{2} d^\top J_k^\top J_k d, \; \text{s.t.} \|d\| \le \Delta_k. \tag{2.10}$$

   The minimization of the objective function without the trust region constraint $\|d\| \le \Delta_k$ leads to the Gauss-Newton method. The addition of the trust-region ensures the global convergence.

4. Quasi-Newton method: $x_{k+1} = x_k + d_k$, where d_k is the solution of

$$\min_{d \in \mathbb{R}^n} \; d^\top \nabla f(x_k) + \frac{1}{2} d^\top (J_k^\top J_k + B_k) d, \; \text{s.t.} \|d\| \le \Delta_k. \tag{2.11}$$

   where B_k is a Quasi-Newton approximation to the term $\sum_{j=1}^{m} r_j(x) \nabla^2 r_j(x)$.

Clearly, the Hessian (2.6) of the KS energy functional is also split in two parts. In particular, one part of the Hessian is defined by $H(X)$ which is already available in the gradient. Hence, using the relationship (2.8), SCF can be viewed as a counterpart of the Gauss-Newton method without considering the "complicate" part of the Hessian. The rest of the paper shows that adding the trust-region like regularization can indeed guarantee convergence while the (approximate) full Hessian can accelerate convergence.

The superlinear or quadratic convergence of Gauss-Newton or Levenberg-Marquardt methods applies for zero residue problems, since in this case the $J_k^\top J_k$ will asymptotically approximates the true Hessian. However, in our problem, the missing part of the Hessian usually does not vanish at the optimum. Hence, for fast local convergence, an asymptotic accurate approximation of the true full Hessian would be necessary in the algorithm.

3. Regularization by Trust-Region Type Techniques.

3.1. The Regularized SCF Subproblems. We begin with an initial estimate X_0 with $X_0^* X_0 = I$ and present the regularized SCF subproblems at kth iteration. A direct mimicking of the Levenberg-Marquardt

method is

$$\begin{aligned} \min \quad & \frac{1}{2}\langle H(X_k)X, X\rangle \\ \text{s.t.} \quad & X^*X = I, \\ & \|X - X_k\|_F \le \Delta_k. \end{aligned} \tag{3.1}$$

Unfortunately, it is not clear whether (3.1) can be solved efficiently with the additional trust-region constraint. Yang, Meza and Wang replaced the term $\|X - X_k\|_F^2$ by $\|XX^* - X_kX_k^*\|_F^2$ in [27] and penalized it in the objective function. Since both X and X_k are orthogonal matrices, the quadratic terms in $\|XX^* - X_kX_k^*\|_F^2$ are constant. Then, the subproblem becomes

$$\begin{aligned} \min \quad & \frac{1}{2}\langle H(X_k)X, X\rangle + \frac{\tau_k}{4}\|XX^* - X_kX_k^*\|_F^2 \equiv \frac{1}{2}\langle H(X_k)X, X\rangle - \frac{\tau_k}{2}\mathrm{tr}(X^*X_kX_k^*X) \\ \text{s.t.} \quad & X^*X = I, \end{aligned} \tag{3.2}$$

which leads to a standard linear eigenvalue problem. However, their numerical experiments show that solving a sequence of the trust region subproblems themselves is not necessarily faster than SCF. Therefore, they embedded the trust-region method in a direct constrained minimization algorithm.

Motivated by the Levenberg-Marquardt method, our method penalizes the distance $\|X - X_k\|_F$ in the objective function:

$$m_k^L(X) := \Re\langle H(X_k)X_k, X - X_k\rangle + \frac{1}{2}\Re\langle H(X_k)(X - X_k), X - X_k\rangle + \frac{\tau_k}{\nu}\|X - X_k\|_F^\nu, \tag{3.3}$$

where ν is either 2 or 3. The regularization parameter $\tau_k > 0$ plays a similar role as the trust-region radius Δ_k in (2.10) and they act reciprocal to each other in the sense that increasing τ_k corresponds to decreasing Δ_k and vice versa. A special choice of τ_k will be discussed in sub-section 3.2. Then the regularized SCF subproblem is

$$\begin{aligned} \min \quad & m_k^L(X) \\ \text{s.t.} \quad & X^*X = I. \end{aligned} \tag{3.4}$$

Suppose $\nu = 2$ and let Z_k denote the optimal solution of (3.4). Then the first-order optimality conditions of (3.4) are simply

$$(H(X_k) + \tau_k I)\, Z_k = Z_k\Lambda_{k+1} + \tau_k X_k \text{ and } Z_k^*Z_k = I, \tag{3.5}$$

which seems to be no more difficult than solving the linear eigenvalue problem except there is an extra linear term in the first equation of (3.5). When $\nu = 3$, problem (3.4) corresponds to the cubic regularization used in [4] which has been proved to be useful in nonlinear programming.

Since the term $\frac{1}{2}\langle H(X_k)X, X\rangle$ can be far away from the exact curvature, in particular, on difficult problems, it is often helpful to consider using the exact Hessian when solving (3.4). Combining Lemma 2.1

and (3.4), we obtain an approximate Taylor expansion to $E(X)$ as

$$
\begin{aligned}
(3.6)\quad m_k^N(X) := & \Re\langle H(X_k)X_k, X - X_k\rangle + \frac{1}{2}\Re\langle H(X_k)(X - X_k), X - X_k\rangle \\
& + \frac{1}{2}\Re\left\langle X - X_k, \mathrm{diag}(J((\bar{X}_k \odot (X - X_k) + X_k \odot (\bar{X} - \bar{X}_k))e)X_k\right\rangle + \frac{\tau_k}{\nu}\|X - X_k\|_F^\nu,
\end{aligned}
$$

where $J = \Re L^\dagger + \frac{\partial^2 \epsilon_{xc}}{\partial \rho^2} e$, and then compute the Newton's step:

$$
(3.7)\qquad \begin{aligned} \min \quad & m_k^N(X) \\ \text{s.t.} \quad & X^*X = I. \end{aligned}
$$

When the computational cost of the second order derivative $\frac{\partial^2 \epsilon_{xc}}{\partial \rho^2}$ of the exchange correlation energy is expensive, a quasi-Newton approximation B_k can be used to substitute $\frac{\partial^2 \epsilon_{xc}}{\partial \rho^2}$ at iteration k. Specifically, let $\widehat{J} = \Re L^\dagger + B_k$ and construct the approximation:

$$
\begin{aligned}
(3.8)\quad m_k^Q(X) := & \Re\langle H(X_k)X_k, X - X_k\rangle + \frac{1}{2}\Re\langle H(X_k)(X - X_k), X - X_k\rangle \\
& + \frac{1}{2}\Re\left\langle X - X_k, \mathrm{diag}(\widehat{J}((\bar{X}_k \odot (X - X_k) + X_k \odot (\bar{X} - \bar{X}_k))e)X_k\right\rangle + \frac{\tau_k}{\nu}\|X - X_k\|_F^\nu.
\end{aligned}
$$

Then the subproblem involving the quasi-Newton approximation is

$$
(3.9)\qquad \begin{aligned} \min \quad & m_k^Q(X) \\ \text{s.t.} \quad & X^*X = I. \end{aligned}
$$

The regularized SCF subproblems (3.4), (3.7) and (3.9) can be solved by an efficient feasible method proposed in [23]. Let $m_k(X)$ be any of $m_k^L(X)$, $m_k^N(X)$ or $m_k^Q(X)$. Starting from $X_k^{(0)} = X_k$, the method generates the $(i+1)$-th iteration by the scheme:

$$
(3.10)\qquad X_k^{(i+1)} := Y_k^{(i)}(\sigma) = \left(I - \frac{\sigma}{2}A^{(i)}\right)^{-1}\left(I + \frac{\sigma}{2}A^{(i)}\right)X_k^{(i)},
$$

where σ is a suitable step size and $A^{(i)}$ is a skew-symmetric matrix defined by

$$
A^{(i)} = X_k^{(i)}\left(\nabla m_k(X_k^{(i)})\right)^* - \left(\nabla m_k(X_k^{(i)})\right)\left(X_k^{(i)}\right)^*.
$$

The matrix $A^{(i)}$ has rank $2p$. In many cases, p is much smaller than $n/2$. It follows from the Sherman-Morrison-Woodbury (SMW) theorem that one only needs to invert a smaller $2p \times 2p$ matrix. The total flops for computing $Y_k^{(i)}(\sigma)$ is $8np^2 + O(p^3)$, and updating $Y_k^{(i)}(\sigma)$ for a different σ needs $4np^2 + O(p^3)$ flops. We outline the major steps of the basic version of the algorithm in Algorithm 1. For more details we refer the reader to [23].

Algorithm 1: A Curvilinear Search Method for Solving the Regularized SCF Subproblems

Input $X_k^{(0)} = X_k$. Set $i = 0$.

while *stopping conditions not met* **do**

$A^{(i)} \leftarrow X_k^{(i)} \left(\nabla m_k(X_k^{(i)})\right)^* - \left(\nabla m_k(X_k^{(i)})\right) \left(X_k^{(i)}\right)^*$.

Compute a suitable step size σ.

$X_k^{(i+1)} \leftarrow Y_k^{(i)}(\sigma) = \left(I - \frac{\sigma}{2}A^{(i)}\right)^{-1} \left(I + \frac{\sigma}{2}A^{(i)}\right) X_k^{(i)}$.

$i \leftarrow i + 1$.

We should point out that the feasible point method in [23] can also be applied to solve (2.2) directly. Although the algorithm can be competitive to SCF on many problems, its convergence is often slowed down when the iterate is close to the optimal solution, and thus it can take a lot of iterations to obtain a high accurate solution. Usually, fast local convergence cannot be expected if only the gradient information is used, in particular, for difficult non-quadratic problems. However, the method is ideal for solving the regularized SCF subproblems (3.4), (3.7) and (3.9) since i) the computational cost of the gradient of these subproblems is usually cheaper than that of (2.2); ii) it is not necessary to solve these subproblems to a high accuracy, especially, at the early stage of the algorithm when a good starting guess is not available.

3.2. The Regularized SCF Framework. We next present the regularized SCF framework starting from a feasible initial point X_0 and the regularization parameter τ_0 using a fixed regularized SCF subproblem either (3.4) or (3.7) or (3.9) for all iterations. At the k-th iteration, an optimal or approximate optimal solution Z_k is generated as follows. Let $Z_k(\sigma)$ denote the feasible curve

$$Z_k(\sigma) := \left(I + \frac{\sigma}{2}W_k\right)^{-1} \left(I - \frac{\sigma}{2}W_k\right) X_k, \tag{3.11}$$

where

$$W_k = G_k X_k^* - X_k G_k^*$$

with $G_k = \nabla E(X_k) = H(X_k)X_k$. An approximate Cauchy point is defined as $X_k^c := Z_k(\sigma_k^c)$, where σ_k^c is the maximal value of $\{1, \beta, \beta^2, \beta^3, ...\} \subset (0, 1]$ such that a descent condition in the model function m_k is guaranteed:

$$m_k(X_k^c) \leq -\frac{1}{2}\gamma \cdot \sigma_k^c \|W_k\|_F^2, \tag{3.12}$$

where $\beta, \gamma \in (0, 1)$. Then, for $\alpha \in (0, 1)$, the trial point Z_k should satisfy a fraction of Cauchy decrease condition

$$m_k(Z_k) \leq \alpha \cdot m_k(X_k^c) \text{ and } \quad Z_k^* Z_k = I. \tag{3.13}$$

Note that Algorithm 1 can be applied to obtain Z_k.

Generally speaking, an algorithm cannot be guaranteed to converge globally if X_{k+1} is set directly to the trial point Z_k obtained from a model with a fixed τ_k. In order to decide whether the trial point Z_k should be accepted and whether the regularization parameter should be updated or not, we calculate the

Algorithm 2: The Regularized SCF Method

S0 Initialize: Given a feasible initial solution X_0 with $X_0^* X_0 = I$ and initial regularization parameter $\tau_0 > 0$. Choose $0 < \eta_1 \le \eta_2 < 1$, $1 < \gamma_1 \le \gamma_2$ and $\alpha, \beta, \gamma \in (0,1)$. Set iteration $k := 0$.
Choose one of the subproblems (3.4) or (3.7) or (3.9) and set $\nu = 2$ or 3.
while *stopping conditions not met* **do**
S1 Compute a new trial point Z_k satisfying the approximate Cauchy condition (3.13).
S2 Compute the ratio ρ_k via (3.14).
S3 Update X_{k+1} from the trial point Z_k based on (3.15).
S4 Update τ_k according to (3.16).
$k \leftarrow k + 1$.

ratio between the actual reduction of the objective function $E(X)$ and predicted reduction:

$$\rho_k = \frac{E(Z_k) - E(X_k)}{m_k(Z_k)}. \tag{3.14}$$

If $\rho_k \ge \eta_1 > 0$, then the iteration is successful and we set $X_{k+1} = Z_k$; otherwise, the iteration is not successful and we set $X_{k+1} = X_k$, that is,

$$X_{k+1} = \begin{cases} Z_k, & \text{if } \rho_k \ge \eta_1, \\ X_k, & \text{otherwise.} \end{cases} \tag{3.15}$$

Then the regularization parameter τ_{k+1} is updated as

$$\tau_{k+1} \in \begin{cases} (0, \tau_k] & \text{if } \rho_k > \eta_2, \\ [\tau_k, \gamma_1 \tau_k] & \text{if } \eta_1 \le \rho_k \le \eta_2, \\ [\gamma_1 \tau_k, \gamma_2 \tau_k] & \text{otherwise.} \end{cases} \tag{3.16}$$

where $0 < \eta_1 \le \eta_2 < 1$ and $1 < \gamma_1 \le \gamma_2$. These parameters determine how aggressively the regularization parameter is decreased when an iteration is successful or it is increased when an iteration is unsuccessful. In practice, the performance of regularized SCF algorithm is not very sensitive to the values of the parameters.

The complete regularized SCF algorithm to solve (2.2) is summarized in Algorithm 2.

4. Convergence Theory. We now analyze the convergence of Algorithm 2 based on either the model (3.3), (3.6) or (3.8) with the cubic regularization, i.e., $\nu = 3$. The analysis can be similarly extended to the algorithm using quadratically regularized SCF subproblems as well. To allow an uniform treatment of the convergence analysis, we substitute the respective second order parts of the different models and study the generalized model

$$\begin{aligned} \min_{Z_k} \quad & m_k(Z_k) := \Re \langle H(X_k) X_k, Z_k - X_k \rangle + b_k(Z_k - X_k, Z_k - X_k) + \frac{\tau_k}{3} \|Z_k - X_k\|_F^3 \\ \text{s.t.} \quad & Z_k^* Z_k = I, \end{aligned} \tag{4.1}$$

where $b_k : \mathbb{C}^{n\times p} \times \mathbb{C}^{n\times p} \to \mathbb{R}$ is a symmetric R-bilinear form

$$b_k(\lambda X + \mu Y, Z) = \lambda b_k(X, Z) + \mu b_k(Y, Z) \ \text{ and } \ b_k(X, Y) = b_k(Y, X), \qquad \forall\, \lambda, \mu \in \mathbb{R}, \forall\, X, Y, Z \in \mathbb{C}^{n\times p}.$$

In particular, the following specific forms are considered in our model:

$$b_k^L(Y,Z) := \frac{1}{2}\Re\left\langle H(X_k)Y, Z\right\rangle, \;\; b_k^N(Y,Z) := b_k^L(Y,Z) + \frac{1}{2}\Re\left\langle [J(Z\odot \bar{X}_k + X_k \odot \bar{Z})ee^\top]\odot X_k, Y\right\rangle$$

and

$$b_k^Q(Y,Z) := b_k^L(Y,Z) + \frac{1}{2}\Re\left\langle [\widehat{J}(Z\odot \bar{X}_k + X_k \odot \bar{Z})ee^\top]\odot X_k, Y\right\rangle.$$

It is easy to verify, that the above functions b_k^L, b_k^N and b_k^Q satisfy both symmetry and R-linearity, and thus can serve as examplary realizations of b_k.

We first state a few important properties of the derivatives of $Z_k(\sigma)$ with respect σ. It can be verified that

$$Z_k'(\sigma) = -\Big(I + \frac{\sigma}{2}W_k\Big)^{-1} W_k\left(\frac{X_k + Z_k(\sigma)}{2}\right), \tag{4.2}$$

$$Z_k''(\sigma) = -\Big(I + \frac{\sigma}{2}W_k\Big)^{-1} W_k Z_k'(\sigma), \tag{4.3}$$

$$Z_k'''(\sigma) = -\frac{3}{2}\Big(I + \frac{\sigma}{2}W_k\Big)^{-1} W_k Z_k''(\sigma), \tag{4.4}$$

which implies that

$$Z_k(0) = X_k, \qquad Z_k'(0) = -W_kX_k, \text{ and } Z_k''(0) = W_k^2X_k. \tag{4.5}$$

Let $Z, D_1, D_2, D_3 \in \mathbb{C}^{n\times p}$ and set $S := Z - X_k$, then a short calculation shows that

$$\frac{d}{dZ}m_k(Z)[D_1] = \Re\langle G_k, D_1\rangle + 2b_k(S, D_1) + \tau_k\|S\|_F \cdot \Re\langle S, D_1\rangle,$$

and the second and third order derivative of the model function m_k are given by

$$\begin{aligned}
\frac{d^2}{d^2Z}m_k(Z)[D_1,D_2] &= 2b_k(D_1,D_2) + \tau_k\|S\|_F^{-1}\cdot\Re\left\langle S, D_1\right\rangle\Re\left\langle S, D_2\right\rangle + \tau_k\|S\|_F\cdot\Re\left\langle D_1, D_2\right\rangle,\\
\frac{d^3}{d^3Z}m_k(Z)[D_1,D_2,D_3] &= -\tau_k\|S\|_F^{-3}\cdot\Re\left\langle S, D_1\right\rangle\Re\left\langle S, D_2\right\rangle\Re\left\langle S, D_3\right\rangle + \tau_k\|S\|_F^{-1}\cdot\Re\left\langle S, D_1\right\rangle\Re\left\langle D_2, D_3\right\rangle\\
&\quad +\tau_k\|S\|_F^{-1}\cdot\Re\left\langle S, D_2\right\rangle\Re\left\langle D_1, D_3\right\rangle + \tau_k\|S\|_F^{-1}\cdot\Re\left\langle S, D_3\right\rangle\Re\left\langle D_1, D_2\right\rangle.
\end{aligned}$$

Hence, using the invariance of the trace operator under cyclic permutation, we obtain

$$\begin{aligned}
\frac{\partial}{\partial\sigma}m_k(Z_k(\sigma))\Big|_{\sigma=0} &= \frac{d}{dZ}m_k(Z_k(0))[Z_k'(0)] = \frac{d}{dZ}m_k(X_k)[-W_kX_k] = -\Re\operatorname{tr}(G_k^*W_kX_k)\\
&= -\frac{1}{2}\Re\operatorname{tr}(X_kG_k^*G_kX_k^* - G_kX_k^*G_kX_k^* - X_kG_k^*X_kG_k^* + G_kX_k^*X_kG_k^*) = -\frac{1}{2}\|W_k\|_F^2.
\end{aligned} \tag{4.6}$$

4.1. Analysis of the Cauchy point. We assume:

ASSUMPTION 4.1. *Suppose that the R-bilinear form b_k is bounded with constant κ_B, i.e.,*

$$|b_k(X,Y)| \le \kappa_B\|X\|_F\|Y\|_F, \qquad \forall\, X, Y \in \mathbb{C}^{n\times p},\ \forall\, k\in\mathbb{N}$$

and that the matrix $H(X)$ in $\nabla E(X) = H(X)X$ is Lipschitz continuous on the convex hull $\operatorname{conv}(\mathcal{M}_n^p)$ *of the Stiefel manifold with constant $L \in \mathbb{R}$, i.e.,*

$$\|H(X) - H(Y)\|_F \leq L\|X - Y\|_F, \qquad \forall\, X, Y \in \operatorname{conv}(\mathcal{M}_n^p).$$

Let Assumption 4.1 hold, then the continuity of $H(X)$ and the compactness of the Stiefel manifold $\mathcal{M}_n^p := \{X \in \mathbb{C}^{n\times p} \mid X^*X = I\}$ (see, for example, Chapter 3.3.2 in [1]) immediately imply the boundedness of $H(X)$ on $\mathcal{M}_n^p$ and $\operatorname{conv}(\mathcal{M}_n^p)$. In particular, there exist $\kappa_H, \kappa_C > 1$, such that

$$\|H(X)\|_F \leq \kappa_H \quad \forall X \in \mathcal{M}_n^p \quad \text{and} \quad \|H(X)\|_F \leq \kappa_C \quad \forall X \in \operatorname{conv}(\mathcal{M}_n^p).$$

LEMMA 4.2. *Suppose there exists an open set $\Omega \supset \Omega_0 := \rho(\operatorname{conv}(\mathcal{M}_n^p))$, such that*

(i) ϵ_{xc} is differentiable on Ω and its derivative μ_{xc} is Lipschitz continuous on Ω_0. Then, Assumption 4.1 is satisfied for $b_k \equiv b_k^L$.
(ii) ϵ_{xc} is twice differentiable on Ω and its second derivative $\frac{\partial^2}{\partial^2\rho}\epsilon_{xc}$ is continuous on Ω_0 or the quasi-Newton approximation $\{\|B_k\|\}$ is uniformly bounded. Then, Assumption 4.1 is satisfied for $b_k \equiv b_k^N$ or $b_k \equiv b_k^Q$.

Proof. Due to the Lipschitz continuity of μ_{xc} on Ω_0 and of ρ on $\operatorname{conv}(\mathcal{M}_n^p)$, $H(X)$ is Lipschitz continuous on $\operatorname{conv}(\mathcal{M}_n^p)$. In case *(ii)* the Lipschitz continuity of μ_{xc} is implied by the continuity of $\frac{\partial^2}{\partial^2\rho}\epsilon_{xc}$ on the compact set Ω_0. The boundedness of the R-bilinear forms follows from the continuity of the respective derivatives on $\rho(\operatorname{conv}(\mathcal{M}_n^p))$. ∎

We start our analysis with the following Armijo-type result:

LEMMA 4.3. *Let Assumption 4.1 hold and let $\gamma \in (0,1)$ be arbitrary. Then there exists a constant $\zeta = \zeta(\tau_k, \|W_k\|_F) > 0$ that only depends on the regularization parameter τ_k and $\|W_k\|_F$ with*

$$m_k(Z_k(\sigma)) \leq -\frac{1}{2}\gamma \cdot \sigma\|W_k\|_F^2 \quad \text{for all } \sigma \in [0, \zeta]. \tag{4.7}$$

In particular, it holds with $\zeta(\tau_k, \|W_k\|_F) = C \cdot \min\left\{1, \frac{1}{\sqrt{(1+\tau_k)\|W_k\|_F}}\right\}$, where $C := \frac{1-\gamma}{11np(\kappa_H+\kappa_B)}$.

Proof. Apparently $\sigma = 0$ fulfils (4.7). Assume that $\sigma > 0$ is sufficiently small. Then we obtain for

arbitrary but fixed $k \in \mathbb{N}$ that

$$
\begin{aligned}
&\frac{m_k(Z_k(\sigma))}{\sigma} + \frac{1}{2}\gamma \cdot \|W_k\|_F^2 \\
&= -\frac{1}{2}(1-\gamma)\cdot\|W_k\|_F^2 + \frac{1}{2}\sigma \cdot \frac{\partial^2}{\partial \tau^2} m_k(Z_k(\tau))\Big|_{\tau=0} + \frac{1}{6}\sigma^2 \cdot \frac{\partial^3}{\partial \tau^3} m_k(Z_k(\tau))\Big|_{\tau=\xi\sigma} \\
&= -\frac{1}{2}(1-\gamma)\cdot\|W_k\|_F^2 + \frac{1}{2}\sigma \cdot \left\{ \frac{d^2}{d^2 Z} m_k(Z_k(\tau))[Z_k'(\tau), Z_k'(\tau)] + \frac{d}{dZ} m_k(Z_k(\tau))[Z_k''(\tau)] \right\}_{\tau=0} \\
&\quad + \frac{1}{6}\sigma^2 \cdot \left\{ \frac{d^3}{d^3 Z} m_k(Z_k(\tau))[Z_k'(\tau), Z_k'(\tau), Z_k'(\tau)] + 3\frac{d^2}{d^2 Z} m_k(Z_k(\tau))[Z_k'(\tau), Z_k''(\tau)] \right. \\
&\quad \left. + \frac{d}{dZ} m_k(Z_k(\tau)[Z_k'''(\tau)] \right\}_{\tau=\xi\sigma} \\
&\leq -\frac{1}{2}(1-\gamma)\cdot\|W_k\|_F^2 + \sigma p \cdot \left(\kappa_B + \frac{1}{2}\kappa_H\right)\cdot\|W_k\|_F^2 + \frac{1}{6}\sigma^2 \cdot \left\{ \frac{d^3}{d^3 Z} m_k(Z_k(\tau))[Z_k'(\tau), Z_k'(\tau), Z_k'(\tau)] \right. \\
&\quad \left. + 3\frac{d^2}{d^2 Z} m_k(Z_k(\tau))[Z_k'(\tau), Z_k''(\tau)] + \frac{d}{dZ} m_k(Z_k(\tau))[Z_k'''(\tau)] \right\}_{\tau=\xi\sigma},
\end{aligned}
$$

where we used a Taylor expansion with suitable $\xi \in [0,1]$, (4.5) and (4.6). Due to the construction of the matrix $Z_k(\tau)$, it is orthogonal for all $k \in \mathbb{N}$ and $\tau \geq 0$, i.e., we have

$$
\|X_k\|_F = \mathrm{tr}(X_k^* X_k)^{\frac{1}{2}} = \sqrt{p} \qquad \|Z_k(\tau)\|_F = \mathrm{tr}(Z_k(\tau)^* Z_k(\tau))^{\frac{1}{2}} = \sqrt{p}. \tag{4.8}
$$

Furthermore, it holds

$$
\|W_k\|_F = \|H(X_k)X_k X_k^* - X_k X_k^* H(X_k)\|_F \leq 2\kappa_H \mathrm{tr}(X_k X_k^*)^{\frac{1}{2}} = 2\kappa_H\sqrt{p}. \tag{4.9}
$$

Now, assuming $\sigma < \frac{1}{2\kappa_H\sqrt{p}}$ and applying the properties of the Neumann series, we obtain

$$
\frac{\sigma}{2}\|W_k\|_F \leq \sigma\cdot\kappa_H\sqrt{p} < \frac{1}{2} \quad \Rightarrow \quad \left\| \left(I + \frac{\sigma}{2}W_k\right)^{-1} \right\|_F < \|I_{n\times n}\|_F - 1 + \frac{1}{1-\frac{\sigma}{2}\|W_k\|_F} \leq \sqrt{n} + 1 \leq 2\sqrt{n}.
$$

Hence, for all $\tau \leq \sigma$, it follows that

$$
\|Z_k(\tau) - X_k\|_F \leq \|Z_k(\tau)\|_F + \|X_k\|_F = 2\sqrt{p}, \tag{4.10}
$$

$$
\|Z_k'(\tau)\|_F \leq \frac{1}{2}\left\| \left(I + \frac{\tau}{2}W_k\right)^{-1} \right\|_F \cdot \|W_k\|_F \cdot (\|X_k\|_F + \|Z_k(\tau)\|_F) \leq 2\sqrt{np}\cdot\|W_k\|_F, \tag{4.11}
$$

$$
\|Z_k''(\tau)\|_F \leq 4n\sqrt{p}\cdot\|W_k\|_F^2, \tag{4.12}
$$

$$
\|Z_k'''(\tau)\|_F \leq 12n\sqrt{np}\cdot\|W_k\|_F^3. \tag{4.13}
$$

Using (4.8)–(4.13) and $\xi \in [0,1]$ yields

$$
\begin{aligned}
\left| \frac{d}{dZ} m_k(Z_k(\xi\sigma))[Z_k'''(\xi\sigma)] \right| &\leq (\|G_k\|_F + 2\kappa_B\|Z_k(\xi\sigma) - X_k\|_F + \tau_k\|Z_k(\xi\sigma) - X_k\|_F^2)\cdot\|Z_k'''(\xi\sigma)\|_F \\
&\leq 12np\sqrt{n}(\kappa_H + 4\kappa_B + 4\tau_k\sqrt{p})\|W_k\|_F^3,
\end{aligned} \tag{4.14}
$$

$$\left|\frac{d^2}{d^2Z}m_k(Z_k(\xi\sigma))[Z_k'(\xi\sigma),Z_k''(\xi\sigma)]\right| \le (2\kappa_B + 2\tau_k\|Z_k(\xi\sigma)-X_k\|_F)\cdot\|Z_k'(\xi\sigma)\|_F\|Z_k''(\xi\sigma)\|_F$$
$$\le 16np\sqrt{n}(\kappa_B+2\tau_k\sqrt{p})\cdot\|W_k\|_F^3, \tag{4.15}$$

and

$$\left|\frac{d^3}{d^3X}m_k(Z_k(\xi\sigma))[Z_k'(\xi\sigma),Z_k'(\xi\sigma),Z_k'(\xi\sigma)]\right| \le 4\tau_k\cdot\|Z_k'(\xi\sigma)\|_F^3 \le 32np\sqrt{np}\tau_k\cdot\|W_k\|_F^3. \tag{4.16}$$

Applying (4.14), (4.15) and (4.16), we obtain

$$\frac{\partial^3}{\partial\tau^3}m_k(Z_k(\tau))\Big|_{\tau=\xi\sigma} \le 176np\sqrt{np}(\kappa_H+\kappa_B)\cdot(1+\tau_k)\cdot\|W_k\|_F^3.$$

Thus, for all $\sigma < \frac{1}{2\kappa_H\sqrt{p}}$, it holds

$$\frac{m_k(Z_k(\sigma))}{\sigma}+\frac{1}{2}\gamma\cdot\|W_k\|_F^2 \le -\frac{1}{2}(1-\gamma)\|W_k\|_F^2+\sigma\cdot p(\kappa_B+\kappa_H)\|W_k\|_F^2+\sigma^2\cdot\frac{176}{6}np\sqrt{np}(\kappa_H+\kappa_B)\cdot(1+\tau_k)\|W_k\|_F^3.$$

Then, the Armijo condition (4.7) is satisfied by setting

$$\sigma\in[0,\zeta],\quad \zeta(\tau_k,\|W_k\|_F) := \min\left\{\frac{1-\gamma}{11np(\kappa_H+\kappa_B)},\frac{1-\gamma}{11np(\kappa_H+\kappa_B)\sqrt{(1+\tau_k)\|W_k\|_F}}\right\}. \tag{4.17}$$

□

Lemma 4.3 shows that the Armijo condition is satisfied for $\sigma_k^c \le \zeta(\tau_k,\|W_k\|_F)$, i.e. Condition (3.12) is well-defined.

COROLLARY 4.4. *Let Assumption 4.1 hold and suppose that the trial point Z_k satisfies* (3.13)*. Then for all $k\ge 0$ we have*

$$-m_k(Z_k) \ge -\alpha\cdot m_k(X_k^c) \ge \frac{1}{2}\alpha\beta\gamma C\cdot\|W_k\|_F\cdot\min\left\{\|W_k\|_F,\sqrt{\frac{\|W_k\|_F}{1+\tau_k}}\right\}. \tag{4.18}$$

Proof. By using Lemma 4.3, we have $\sigma_k^c \ge \beta\zeta(\tau_k,\|W_k\|_F)$. Then, it follows from (3.12) that

$$m_k(X_k^c) \le -\frac{1}{2}\gamma\sigma_k^c\|W_k\|_F^2 \le -\frac{1}{2}\beta\gamma C\cdot\min\left\{1,\frac{1}{\sqrt{(1+\tau_k)\|W_k\|_F}}\right\}\|W_k\|_F^2.$$

This immediately implies (4.18). □

4.2. Global Convergence. Motivated by the similarity of the previous subsection and Lemma 2.1 in [4], we follow the ideas in [4] and extend the existing theory to problems with orthogonality constraints.

LEMMA 4.5. *Let Assumption 4.1 hold and suppose that Z_k satisfies* (3.13)*. Then*

$$\|Z_k-X_k\|_F \le \frac{5(\kappa_H+\kappa_B)}{\tau_k}\max\{1,\sqrt{\tau_k p}\}. \tag{4.19}$$

Proof. Using Corollary 4.4, $\kappa_H, p \geq 1$ and following the proof of Lemma 2.2 in [4], we obtain

$$\|Z_k - X_k\|_F \leq \frac{3}{\tau_k} \max\{(3\kappa_B)/2, \sqrt{\tau_k \|G_k\|_F}\} \leq \frac{5(\kappa_H + \kappa_B)}{\tau_k} \max\{1, \sqrt{\tau_k p}\}.$$

□

LEMMA 4.6. *Suppose that Assumption 4.1 holds and that Z_k satisfies* (3.13)*. Suppose furthermore that $W_k \neq 0$ and that*

$$\sqrt{\tau_k} \geq \max\left\{\frac{1}{\sqrt{p}}, \frac{\sqrt{1+p} \cdot \kappa}{(1-\eta_2) \cdot \|W_k\|_F \cdot \min\{\|W_k\|_F, \sqrt{\|W_k\|_F}\}}\right\} \geq 0, \tag{4.20}$$

where $\kappa := \frac{50p(\kappa_H+\kappa_B)^2 \cdot (L\sqrt{p}+\kappa_B+\kappa_C)}{\alpha\beta\gamma C}$ and $\gamma \in (0,1)$. Then iteration k is very successful, i.e. $\rho_k \geq \eta_2$ and

$$\tau_{k+1} \leq \tau_k. \tag{4.21}$$

Proof. Since the trial point Z_k satisfies (3.13), it holds

$$-m_k(Z_k) \geq \frac{1}{2}\alpha\beta\gamma C \cdot \|W_k\|_F \cdot \min\left\{\|W_k\|_F, \sqrt{\frac{\|W_k\|_F}{1+\tau_k}}\right\}.$$

Using a Taylor expansion with suitable $\xi \in [0,1]$, we have

$$E(Z_k) = E(X_k) + \Re\left\langle H(X_k + \xi(Z_k - X_k))(X_k + \xi(Z_k - X_k)), Z_k - X_k\right\rangle,$$

which together with Assumption 4.1 and $X_k + \xi(Z_k - X_k) \in \operatorname{conv}(\mathcal{M}_n^p)$ leads to the estimate

$$\begin{aligned} E(Z_k) - E(X_k) - m_k(Z_k) &= \Re\left\langle [H(X_k + \xi(Z_k - X_k)) - H(X_k)]X_k, Z_k - X_k\right\rangle - b_k(Z_k - X_k, Z_k - X_k) \\ &\quad + \xi \cdot \Re\left\langle H(X_k + \xi(Z_k - X_k))(Z_k - X_k), Z_k - X_k\right\rangle - \frac{\tau_k}{3}\|Z_k - X_k\|_F^3 \\ &\leq (L\sqrt{p} + \kappa_B + \kappa_C)\|Z_k - X_k\|_F^2. \end{aligned}$$

Consequently, (4.18), (4.19) and (4.20) yield

$$\begin{aligned} 1 - \rho_k &= \frac{E(Z_k) - E(X_k) - m_k(Z_k)}{-m_k(Z_k)} \\ &\leq \frac{2(L\sqrt{p} + \kappa_B + \kappa_C) \cdot \sqrt{1+\tau_k}}{\alpha\beta\gamma C \cdot \|W_k\|_F \cdot \min\{\sqrt{1+\tau_k} \cdot \|W_k\|_F, \sqrt{\|W_k\|_F}\}} \cdot \|Z_k - X_k\|_F^2 \\ &\leq \frac{\kappa}{\|W_k\|_F \cdot \min\{\|W_k\|_F, \sqrt{\|W_k\|_F}\}} \cdot \frac{\sqrt{1+\tau_k}}{p\tau_k^2} \cdot \max\{1, \tau_k p\} \\ &= \frac{\kappa}{\|W_k\|_F \cdot \min\{\|W_k\|_F, \sqrt{\|W_k\|_F}\}} \cdot \frac{1}{\sqrt{\tau_k}} \cdot \sqrt{1 + \frac{1}{\tau_k}} \\ &< (1-\eta_2) \cdot \frac{1}{\sqrt{1+p}} \cdot \sqrt{1+p} \\ &= 1 - \eta_2. \end{aligned}$$

The above inequality gives $\rho_k > \eta_2$. Hence, it follows from step S4 of Algorithm 2 that $\tau_{k+1} \leq \tau_k$. ☐

LEMMA 4.7. *Suppose that Assumption 4.1 holds and there exists a constant $\epsilon > 0$ such that $\|W_k\|_F \geq \epsilon$ for all $k \in \mathbb{N}$. Then $(\tau_k)_k$ is bounded, i.e. there exists $L_\epsilon \geq 0$ such that*

$$\tau_k \leq L_\epsilon, \quad \textit{for all } k. \tag{4.22}$$

Proof. Assume that the following bound holds for any $k \geq 0$

$$\sqrt{\tau_k} \geq \max\left\{\frac{1}{\sqrt{p}}, \frac{\sqrt{1+p}\cdot\kappa}{(1-\eta_2)\cdot\epsilon\cdot\min\{\epsilon,\sqrt{\epsilon}\}}\right\} =: \kappa_\tau. \tag{4.23}$$

Then Lemma 4.6 implies that iteration k is very successful with $\tau_{k+1} \leq \tau_k$. Thus, when $\tau_0 \leq \gamma_2\kappa_\tau^2$, it follows $\tau_k \leq \gamma_2\kappa_\tau^2$, $k \geq 0$, where the factor γ_2 shall cover the case when τ_k is less than κ_τ^2 and iteration k is not very successful. Setting $L_\epsilon := \max\left\{\tau_0, \gamma_2\kappa_\tau^2\right\}$ gives (4.22). ☐

The following results are based on Lemma 2.4 and Theorem 2.5 in [4]. We want to point out that the convergence theory in [4] is established under slightly weaker assumptions on the gradient of the objective function. In particular, Cartis et al. provide a generalized version of Lemmas 4.6 and 4.7 and prove convergence without explicitly requiring the Lipschitz continuity of the gradient. However, since this theoretical refinement is not of practical relevance for KS energy minimization problems, our convergence analysis focuses on cases with higher regularity.

LEMMA 4.8. *Suppose that Assumption 4.1 holds and there are only finitely many successful iterations. Then $X_k = X_*$ for all sufficiently large k and $W_* = 0$.*

Proof. Let the last successful iteration be indexed by k_0, then, due to the construction of the algorithm, there holds $X_{k_0+1} = X_k = X_*$, for all $k \geq k_0 + 1$. Since all iterations $k \geq k_0 + 1$ are unsuccessful, the regularization parameter τ_k tends to infinity as $k \to \infty$. If $\|W_{k_0+1}\|_F > 0$, then $\|W_k\|_F = \|W_{k_0+1}\|_F > 0$, for all $k \geq k_0 + 1$, and Lemma 4.7 implies that τ_k is bounded above, $k \geq k_0 + 1$. This contradiction completes the proof. ☐

THEOREM 4.9. *Suppose that Assumption 4.1 holds. Then either*

$$W_l = 0 \quad \textit{for some} \quad l \geq 0$$

$$\textit{or } \lim_{k\to\infty} E(X_k) = -\infty$$

$$\textit{or } \liminf_{k\to\infty} \|W_k\|_F = 0.$$

Proof. Due to Lemma 4.8 we only have to consider the case when infinitely many successful iterations occur. Assume that there exists $\epsilon > 0$ such that

$$\|W_k\|_F \geq \epsilon \quad \text{for all} \quad k \geq 0. \tag{4.24}$$

Let $k \in \mathcal{S}$, where

$$\mathcal{S} := \{k \in \mathbb{N} : \text{ iteration } k \text{ is successful or very successful}\}.$$

Then step S1 of Algorithm 2, Lemma 4.3 and (4.22) imply

$$E(X_k) - E(Z_k) \geq \eta_1 \cdot (-m_k(Z_k)) \geq \frac{\eta_1 \alpha \beta \gamma}{2} \cdot \zeta(\tau_k, \|W_k\|_F) \cdot \|W_k\|_F^2 \geq \frac{\eta_1 \alpha \beta \gamma}{2} \cdot \epsilon^2 \cdot \zeta(L_\epsilon, \epsilon) =: \delta_\epsilon.$$

Summing up over all iterates gives

$$E(X_k) - E(X_{k+1}) = \sum_{j=0, j \in \mathcal{S}}^{k} E(X_j) - E(X_{j+1}) \geq |\mathcal{S} \cap \{1, ..., k\}| \cdot \delta_\epsilon. \tag{4.25}$$

Since $\mathcal{S}$ is not finite, $|\mathcal{S} \cap \{1, ..., k\}| \to \infty$ as $k \to \infty$. Inequality (4.25) now implies that $\lim_{k \to \infty} E(X_k) = -\infty$. On the contrary, if $(E(X_k))_{k \in \mathbb{N}}$ is bounded below, then Assumption (4.24) does not hold and so $(\|W_k\|_F)_{k \in \mathbb{N}}$ has a subsequence converging to zero. □

Corollary 4.10. *Suppose that Assumption 4.1 holds. Then either*

$$W_l = 0 \quad \textit{for some} \quad l \geq 0, \tag{4.26}$$

$$\textit{or} \lim_{k \to \infty} E(X_k) = -\infty, \tag{4.27}$$

$$\textit{or} \lim_{k \to \infty} \|W_k\|_F = 0. \tag{4.28}$$

Proof. Following the ideas of Corollary 2.6 [4] we assume that (4.26) and (4.27) do not hold. Thus, suppose that $(E(X_k))_{k \in \mathbb{N}}$ is bounded below and that there exists a subsequence $(k_i)_{i \in \mathbb{N}} \subset \mathcal{S}$ of successful iterates such that

$$\|W_{k_i}\|_F \geq 2\epsilon, \tag{4.29}$$

for some $\epsilon > 0$ and for all $i \in \mathbb{N}$. Since W_k remains constant whenever an unsuccessful iteration occurs only successful iterates need to be considered. Notice, that there are infinitely many successful iterations since we assumed (4.26) does not hold. This assumption also implies that for each index k_i, there is a first successful iteration $l_i > k_i$ such that $\|W_{l_i}\|_F < \epsilon$. Thus $(l_i)_{i \in \mathbb{N}} \subset \mathcal{S}$ and

$$\|W_k\|_F \geq \epsilon, \quad \text{for } k_i \leq k < l_i, \quad \text{and} \quad \|W_{l_i}\|_F < \epsilon. \tag{4.30}$$

Let $\mathcal{K} := \{k \in \mathcal{S} : k_i \leq k < l_i, i \in \mathbb{N}\}$. Since $\mathcal{K} \subset \mathcal{S}$ we obtain

$$E(X_k) - E(X_{k+1}) \geq \eta_1 \cdot (-m_k(Z_k)) \geq \frac{1}{2} \eta_1 \alpha \beta \gamma C \cdot \epsilon \cdot \min \left\{ \epsilon, \sqrt{\frac{\|W_k\|_F}{1 + \tau_k}} \right\} \quad k \in \mathcal{K}, \tag{4.31}$$

where we used the construction of the algorithm, Corollary 4.4 and (4.30). Now, since $(E(X_k))_{k \in \mathbb{N}}$ is monotonically decreasing and bounded from below, it is convergent and the last inequality implies

$$\sqrt{\frac{\|W_k\|_F}{1 + \tau_k}} \to 0, \quad k \in \mathcal{K}, \quad k \to \infty$$

and, due to (4.30),

$$\tau_k \to \infty, \quad k \in \mathcal{K}, \quad k \to \infty.$$

Thus, for $k \in \mathcal{K}$ sufficiently large, (4.31) implies

$$\frac{1}{\sqrt{\tau_k}} \leq \sqrt{\frac{2}{\epsilon}} \cdot \sqrt{\frac{\|W_k\|_F}{1+\tau_k}} \leq \frac{2\sqrt{2}}{\eta_1 \alpha\beta\gamma C \cdot \epsilon\sqrt{\epsilon}} \cdot (E(X_k) - E(X_{k+1})). \tag{4.32}$$

Hence, we have

$$\|X_{l_i} - X_{k_i}\|_F \leq \sum_{j=k_i, j\in\mathcal{K}}^{l_i-1} \|X_j - X_{j+1}\|_F = \sum_{j=k_i, j\in\mathcal{K}}^{l_i-1} \|Z_j - X_j\|_F, \quad \text{for each } l_i \text{ and } k_i. \tag{4.33}$$

Recalling (4.19), it follows

$$\|Z_j - X_j\|_F \leq \frac{5(\kappa_H + \kappa_B)\sqrt{p}}{\sqrt{\tau_j}} \quad \text{for all } j \in \mathcal{K} \text{ sufficiently large.}$$

Next, (4.32) and (4.33) provide

$$\|X_{l_i} - X_{k_i}\|_F \leq 5(\kappa_H + \kappa_B)\sqrt{p} \sum_{j=k_i, j\in\mathcal{K}}^{l_i-1} \frac{1}{\sqrt{\tau_j}} \leq \frac{10\sqrt{2}(\kappa_H + \kappa_B)\sqrt{p}}{\eta_1 \alpha\beta\gamma C \cdot \epsilon\sqrt{\epsilon}} \cdot \{E(X_{k_i}) - E(X_{l_i})\} \tag{4.34}$$

for all k_i, l_i sufficiently large. Since $(E(X_k))_k$ is convergent, $(E(X_{k_i}) - E(X_{l_i}))_{i\in\mathbb{N}}$ converges to zero as $i \to \infty$. Hence, $\|X_{l_i} - X_{k_i}\|_F$ converges to zero as $i \to \infty$ and by continuity of H, $\|W_{k_i} - W_{l_i}\|_F$ also converges to zero. Due to (4.29) and (4.30), we obtain $\|W_{k_i} - W_{l_i}\|_F \geq \|W_{k_i}\|_F - \|W_{l_i}\|_F \geq \epsilon$ which is a contradiction. □

5. Numerical Results. In order to demonstrate the effectiveness of our regularized SCF method, we implemented it based on a MATLAB toolbox KSSOLV [25] and compared it with the SCF iteration, the trust-region enabled direct constrained minimization algorithm (TRDCM, [25]) and the feasible method for optimization with orthogonality methods (OptM, [23]) on nine standard testing problems. For our comparison, we used three versions of the regularized SCF method: TRQ and TRQH, the versions with the inexact and exact Hessian and quadratic regularization, respectively, and TRCH, which uses the exact Hessian and cubic regularization. All codes were implemented in MATLAB. All the experiments were performed on a Dell Precision T5500 workstation with Intel Xenon(R) E5620 CPU at 2.40GHz (×4) and 12GB of memory running Ubuntu 12.04 and MATLAB 2011b.

All methods were terminated if the residual $\|HX - X(X^*HX)\|_F$ is less than 10^{-6}. For fair comparison, a maximal of 100 iterations is set for all methods except that 1000 is used for OptM, whose computational cost is cheap at each iteration compared with other methods. In addition, OptM, TRQ, TRQH and TRCH were also terminated if the relative changes of the two consecutive iterates and their corresponding objective function values are less than 10^{-8} and 10^{-14}, respectively. The linear eigenvalue problems in SCF were solved by a preconditioned LOBPCG [11] with a maximal of 10 iterations. For TRQ, TRQH and TRCH, we set $\eta_1 = 0.01$, $\eta_2 = 0.9$ and $\tau_k = \omega_k\theta_k$, where ω_k is updated by (3.16). For TRCH, $\theta_k = 1$ and for TRQ and TRQH, $\theta_k = 0.1\|HX_k - X_k(X_k^*HX_k))\|_F$. The k-th regularized SCF subproblem was terminated if a

maximal of 50 iterations is reached or the norm of the projected gradient on the Stiefel manifold was less than

$$\max(tol, \min(0.66\tau_k \|X_k^{(i)} - X_k\|_F, 0.01)),$$

where $X_k^{(i)}$ was the i-th iteration of the k-th regularized SCF subproblem and $tol = \max(\min(0.1\|H_k X_k - X_k(X_k^* H_k X_k)\|_F, 0.1), 10^{-6})$.

Let $\tilde{\gamma} = e^2 \left(\frac{3}{\pi}\right)^{1/3}$ and $\phi_i = (4\pi\rho_i/3)^{-1/3}$. The exchange-correlation energy ϵ_{xc} of (2.1) is defined by

$$\epsilon_{xc} = (\epsilon_{ex} + \epsilon_{ec}) \odot \rho, \tag{5.1}$$

where $\epsilon_{ex} = -\frac{3}{4}\tilde{\gamma}\rho^{\frac{1}{3}}$ and

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

It can be verified that

$$\mu_{xc} := \frac{\partial \epsilon_{ex}}{\partial \rho} = -\tilde{\gamma}\rho^{\frac{1}{3}} = \frac{4}{3}\mu_{ex},$$

$$(\mu_{ec})_i := \left(\frac{\partial \epsilon_{ec}}{\partial \rho}\right)_i = \begin{cases} (A_1 - A_3/3) + (2A_2 - A_4)\phi_i/3 + (A_3 + 2/3A_4\phi_i)\log(\phi_i), & \text{if } \phi_i < 1, \\ B_1 \frac{1+7/6B_2\sqrt{\phi_i}+4/3B_3\phi_i}{(1+B_2\sqrt{\phi_i}+B_3\phi_i)^2}, & \text{if } \phi_i \geq 1. \end{cases}$$

Hence, the terms $\frac{\partial \mu_{xc}}{\partial \rho}$ required for the Hessian-matrix products in Lemma 2.1 are

$$\frac{\partial \mu_{ex}}{\partial \rho} = -\tilde{\gamma}\frac{1}{3}\operatorname{diag}(\rho^{-\frac{2}{3}}), \quad \frac{\partial \mu_{ec}}{\partial \rho} = \operatorname{diag}(\hat{\mu}_{ec} \odot \frac{\partial \phi}{\partial \rho}),$$

where $\frac{\partial \phi}{\partial \rho} = -\frac{4\pi}{9}(4\pi\rho/3)^{-4/3}$ and

$$(\hat{\mu}_{ec})_i = \begin{cases} 2/3A_2 + A_4/3 + A_3/\phi_i + 2/3A_4\log(\phi_i), & \text{if } \phi_i < 1, \\ -\frac{(B_1(7B_2^2+8B_3))/12+B_1(5B_2+16B_3^2\phi_i^{3/2}+21B_2B_3\phi_i)/(12\phi_i^{-1/2})}{(1+B_2\sqrt{\phi_i}+B_3\phi_i)^3}, & \text{if } \phi_i \geq 1. \end{cases}$$

The continuous energy functional is discretized by using the planewave discretization scheme in KSSOLV. By imposing periodic boundary condition, the Laplacian operator L can be expressed as a block circulant matrix with circulant blocks that can be decomposed as $L = F^* D_g F$, and D_g is a diagonal matrix. Each column of X corresponds to n components of a three-dimensional (3D) matrix of size $n_1 \times n_2 \times n_3$. The indices of these n components are determined by the diagonal elements of D_g which are larger than certain thresholding value "ecut". More detailed description of KSSOLV is available in [25]. The matrix sizes of the nine testing examples are summarized in Table 5.1. One of the main computational costs in evaluating the total energy functional (2.1), the gradient and the Hessian-matrix products arises from the 3D Fourier and inverse Fourier transformations of size $n_1 \times n_2 \times n_3$ corresponding to X.

A summary of the computational results is presented in Table 5.2, where "$E(X)$", "resi" and "feasi" denote the total energy functional value, the residual $\|HX - X(X^*HX)\|_F$ and the violation of the orthogonality constraints $\|X^*X - I\|_F$ at the computed solution, respectively, "iter" denotes the total number

TABLE 5.1
Problem Information

name	(n_1, n_2, n_3)	n	p
alanine	(64, 48, 64)	12671	18
al	(64, 64, 64)	16879	12
c12h26	(96, 48, 20)	5709	37
ctube661	(115, 115, 15)	12599	48
glutamine	(64, 55, 74)	16517	29
graphene30	(114, 114, 15)	12279	67
pentacene	(80, 55, 160)	44791	51
ptnio	(63, 34, 30)	4069	43
qdot	(32, 32, 32)	2103	8

of iterations of each algorithm, and "cpu" denotes the CPU time measured in seconds. From the table, we can see that SCF failed at problems "al", "graphene30" and "qdot" and TRDCM failed at problems "graphene30", "ptnio" and "qdot". The pure first-order method OptM often works well and is stopped by the relative change rules on most problems. Since it can take many iterations to achieve a high accuracy, we didn't calibrate its termination rules for a more detailed comparison. Without using the exact Hessian matrix, TRQ failed on instances "ctube661", "graphene30" and "ptnio". Adding the extra Hessian term, both TRQH and TRCH were able to solve all problems and their CPU time are quite competitive to those of SCF and TRDCM. It is interesting to note that TRQ can be faster than TRQH and TRCH on "qdot" to achieve almost the same accuracy. This behavior implies that a rule on switching between the inexact and extra Hessian may be helpful. Although orthogonality is lost on a few problems for OptM, TRQ, TRQH and TRCH, a single orthogonalization step can be executed at the end of these algorithms for a correction.

The typical convergence behaviors of all six methods are demonstrated in Figure 5.1 using four testing problems. The performance of TRQH and TRCH was almost the same. Superlinear convergence can be observed on all problems except "qdot". In fact, they can even achieve quadratic convergence if the tolerance is sufficient small when the regularized SCF subproblem is solved. For example, a tolerance of "1e-6" was used to obtain the lines "TRQH, 1e-6" and "TRCH, 1e-6" in Figure 5.2 and these two variants saved two outer iterations. However, they were more computationally expensive than TRQH and TRCH because more inner iterations were executed. Finally, a summary of the averaged number of the inner iterations and function evaluations for solving the regularized SCF subproblems is presented in Table 5.3. Actually, the cpu time of TRQH and TRCH can often be reduced if only 20 or 30 inner iterations are used.

6. Conclusions. Solving the Kohn-Sham equation or minimizing the electronic total energy functional with respect to electron wave functions is a central problem in electronic structure calculations. In this paper, we propose to regularize the widely used SCF iteration by adding an extra proximal term which enables a rigorous convergence proof. Observing that the SCF iteration can be regarded as an "inexact" Gauss-Newton method, we take advantage of the structure of the exact Hessian of the total energy functional without introducing much additional computational cost. The proposed algorithms compare favorably with SCF for the standard testing problems in the KSSOLV toolbox under the Matlab environment. In particular, our new algorithms exhibit quadratic or superlinear convergence on most test problems and can always attain a high accurate solution even on problems for which SCF fails.

Although our feasible point method in [23] works reasonably well on solving the regularized SCF subproblems, they are still the most computational extensive parts of our algorithm. These subproblems cannot be expressed as a standard linear eigenvalue problem because of the tensor structure in the exact Hessian and the addition of a linear term from the regularization function. Further research on, for example, a

TABLE 5.2
Numerical results on total energy minimization

solver	$E(X)$	iter	resi	feasi	cpu	solver	resi	iter	resi	feasi	cpu
alanine											
SCF	-6.11619212e+01	15	2.4e-07	1.1e-14	59	TRDCM	-6.11619212e+01	100	1.6e-05	1.0e-14	311
OptM	-6.11619212e+01	56	5.3e-06	5.5e-14	46	TRQ	-6.11619212e+01	43	8.7e-06	2.9e-14	137
TRQH	-6.11619212e+01	6	7.7e-07	2.1e-13	65	TRCH	-6.11619212e+01	6	9.9e-07	3.6e-14	51
al											
SCF	-1.57889236e+01	100	1.9e-02	1.1e-14	321	TRDCM	-1.58038176e+01	100	1.7e-05	9.1e-14	260
OptM	-1.58038174e+01	1000	3.5e-05	3.2e-11	640	TRQ	-1.58034407e+01	100	9.8e-04	2.0e-13	378
TRQH	-1.58038176e+01	38	7.9e-07	2.6e-10	712	TRCH	-1.58038176e+01	31	9.8e-07	1.7e-10	590
c12h26											
SCF	-8.15360919e+01	14	6.3e-07	1.8e-14	74	TRDCM	-8.15360919e+01	100	3.0e-05	1.5e-14	363
OptM	-8.15360919e+01	64	1.7e-05	4.5e-14	66	TRQ	-8.15360919e+01	56	1.1e-05	3.7e-14	202
TRQH	-8.15360919e+01	6	9.7e-07	8.4e-14	66	TRCH	-8.15360919e+01	7	8.2e-07	9.0e-14	74
ctube661											
SCF	-1.34638432e+02	15	8.0e-07	3.3e-14	278	TRDCM	-1.34638432e+02	100	3.5e-05	2.5e-14	1395
OptM	-1.34638432e+02	68	1.4e-05	6.1e-14	227	TRQ	-1.34638431e+02	40	1.2e-03	4.5e-14	467
TRQH	-1.34638432e+02	6	2.4e-07	1.0e-13	284	TRCH	-1.34638432e+02	6	9.7e-07	1.3e-13	272
glutamine											
SCF	-9.18394252e+01	16	7.3e-07	1.3e-14	214	TRDCM	-9.18394252e+01	24	5.9e-07	1.3e-14	246
OptM	-9.18394252e+01	76	1.1e-05	2.2e-13	194	TRQ	-9.18394252e+01	60	4.8e-06	1.5e-13	696
TRQH	-9.18394252e+01	8	9.8e-07	3.2e-13	307	TRCH	-9.18394252e+01	13	1.2e-06	4.0e-13	337
graphene30											
SCF	-1.73192488e+02	100	2.3e-02	4.0e-14	2786	TRDCM	-1.73593490e+02	100	9.7e-03	4.0e-14	2107
OptM	-1.73595105e+02	274	4.4e-05	8.8e-12	1252	TRQ	-1.73595071e+02	100	8.8e-03	2.3e-13	1749
TRQH	-1.73595105e+02	18	9.8e-07	9.4e-11	2614	TRCH	-1.73595105e+02	12	6.9e-07	1.6e-11	1510
pentacene											
SCF	-1.31890295e+02	17	5.3e-07	3.4e-14	1022	TRDCM	-1.31890295e+02	100	1.3e-05	3.1e-14	3479
OptM	-1.31890295e+02	116	1.0e-05	5.0e-13	1103	TRQ	-1.31890295e+02	58	6.4e-05	2.2e-13	2960
TRQH	-1.31890295e+02	6	1.0e-06	5.4e-13	1110	TRCH	-1.31890295e+02	8	7.9e-07	6.7e-13	1172
ptnio											
SCF	-2.26788843e+02	55	1.0e-06	1.7e-14	238	TRDCM	-2.26788356e+02	100	9.7e-03	2.2e-14	366
OptM	-2.26788843e+02	431	5.6e-05	1.1e-11	304	TRQ	-2.26763585e+02	100	2.1e-01	1.8e-13	325
TRQH	-2.26788843e+02	24	9.1e-07	1.0e-11	471	TRCH	-2.26788843e+02	25	8.1e-07	2.5e-11	492
qdot											
SCF	2.76981786e+01	100	9.0e-02	2.9e-15	31	TRDCM	2.77010208e+01	100	4.1e-02	2.6e-15	24
OptM	2.76998010e+01	1000	4.4e-05	3.2e-09	56	TRQ	2.76949634e+01	76	9.1e-07	2.8e-09	71
TRQH	2.76949635e+01	76	1.0e-06	5.6e-09	151	TRCH	2.76949635e+01	63	9.3e-07	4.8e-09	111

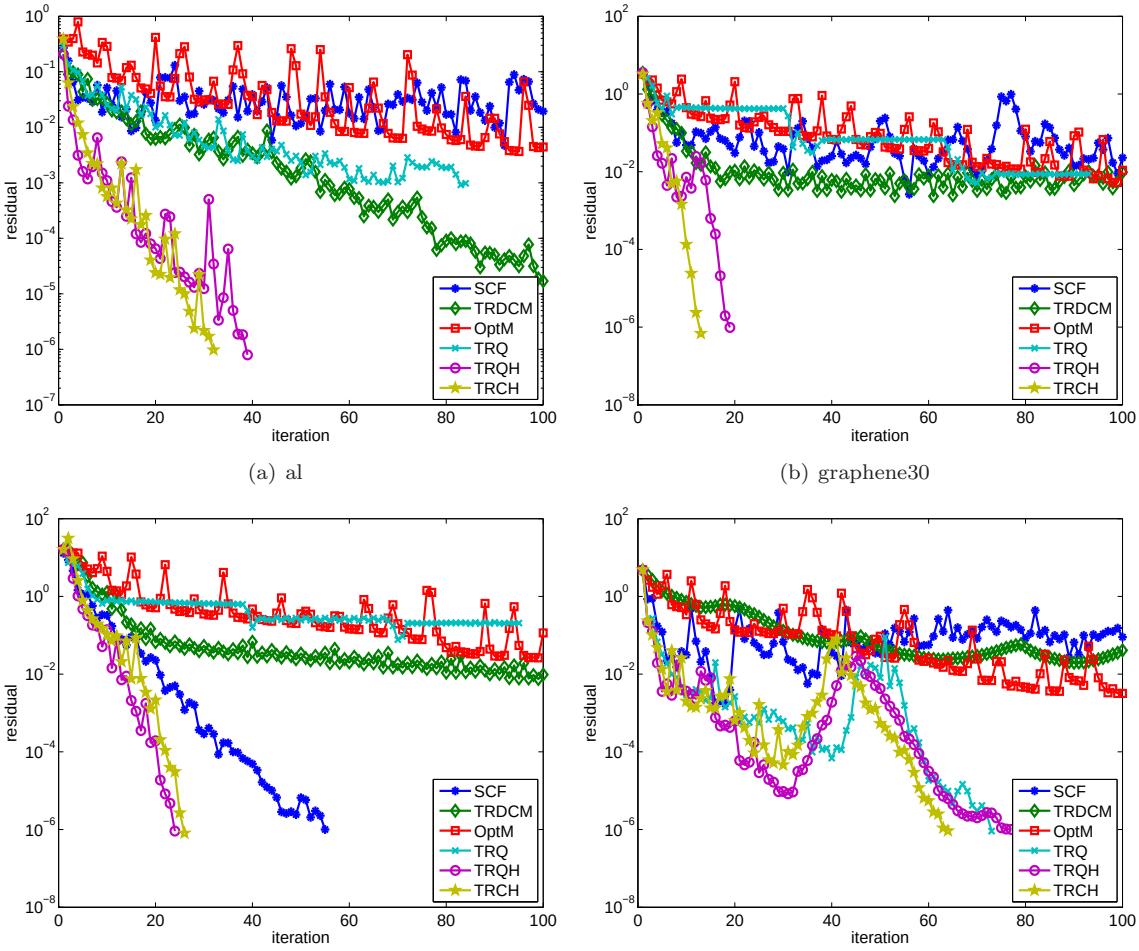
TABLE 5.3
The average of the number of inner iterations and function evaluations for solving the regularized SCF subproblems

	TRQ			TRQH			TRCH		
name	iter	ave.-inner-iter	ave.-nfe	iter	ave.-inner-iter	ave.-nfe	iter	ave.-inner-iter	ave.-nfe
alanine	43	9	11	6	18	21	6	14	16
al	100	14	16	38	44	48	31	44	48
c12h26	56	8	9	6	15	17	7	15	17
ctube661	40	6	8	6	17	20	6	15	19
glutamine	60	10	11	8	19	22	13	13	15
graphene30	100	7	8	18	42	45	12	36	39
pentacene	58	10	12	6	25	28	8	19	22
ptnio	100	9	10	24	35	38	25	35	38
qdot	76	46	50	76	47	50	63	44	47

suitable discretization of the total energy functional, an adaptive rule for switching between using the exact and inexact Hessian in the subproblem, and effective preconditioning techniques are also needed.

Finally, we opine that optimization techniques offer a greater opportunity to further improve the SCF iteration and the simulation of large systems from both theoretical and computational perspectives. Moreover, our adaptive regularization methods can also be naturally applied to solve other optimization problems with orthogonality constraints.

FIG. 5.1. *Residual*



(a) al (b) graphene30

(c) ptnio (d) qdot

Acknowledgements. We would like to thank Xin Liu for the discussions on the trust region methods, Chao Yang for the discussions on the Kohn-Sham equations and KSSOLV.

REFERENCES

[1] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization algorithms on matrix manifolds*, Princeton University Press, Princeton, NJ, 2008.

[2] T. A. Arias, M. C. Payne, and J. D. Joannopoulos, *Ab initio molecular dynamics: Analytically continued energy functionals and insights into iterative solutions*, Phys. Rev. Lett., 69 (1992), pp. 1077–1080.

[3] P. Bendt and A. Zunger, *New approach for solving the density-functional self-consistent-field problem*, Phys. Rev. B, 26 (1982), pp. 3114–3137.

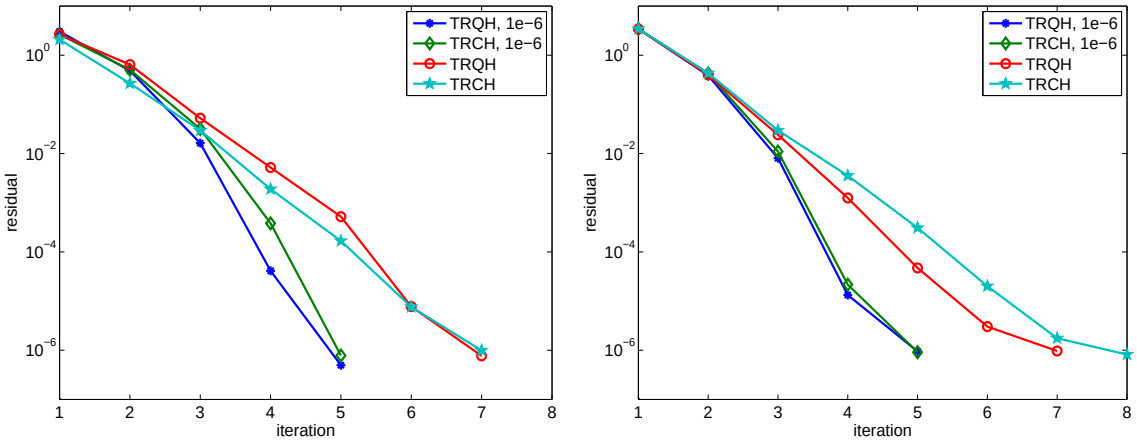
[4] C. Cartis, N. I. M. Gould, and P. L. Toint, *Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results*, Math. Program., 127 (2011), pp. 245–295.

[5] J. B. Francisco, J. M. Martinez, and L. Martinez, *Globally convergent trust-region methods for self-consistent field electronic structure calculations*, The Journal of Chemical Physics, 121 (2004), pp. 10863–10878.

[6] J. B. Francisco, J. M. Martínez, and L. Martínez, *Density-based globally convergent trust-region methods for self-consistent field electronic structure calculations*, J. Math. Chem., 40 (2006), pp. 349–377.

[7] W. Gao, C. Yang, and J. Meza, *Solving a class of nonlinear eigenvalue problems by newtons method*, tech. rep., Lawrence Berkeley National Laboratory, 2009.

FIG. 5.2. *Convergence with respect to the tolerance used in solving the regularized SCF subproblems*



(a) alanine (b) c12h26

[8] W. W. HAGER AND H. ZHANG, *Asymptotic convergence analysis of a new class of proximal point methods*, SIAM J. Control Optim., 46 (2007), pp. 1683–1704.

[9] P. HOHENBERG AND W. KOHN, *Inhomogeneous electron gas*, Phys. Rev., 136 (1964), pp. B864–B871.

[10] G. P. KERKER, *Efficient iteration scheme for self-consistent pseudopotential calculations*, Phys. Rev. B, 23 (1981), pp. 3082–3084.

[11] A. V. KNYAZEV, *Toward the optimal preconditioned eigensolver: locally optimal block preconditioned conjugate gradient method*, SIAM J. Sci. Comput., 23 (2001), pp. 517–541. Copper Mountain Conference (2000).

[12] W. KOHN AND L. J. SHAM, *Self-consistent equations including exchange and correlation effects*, Phys. Rev., 140 (1965), pp. A1133–A1138.

[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] C. LE BRIS, *Computational chemistry from the perspective of numerical analysis*, Acta Numer., 14 (2005), pp. 363–444.

[15] J. NOCEDAL AND S. J. WRIGHT, *Numerical Optimization*, Springer Series in Operations Research and Financial Engineering, Springer, New York, second ed., 2006.

[16] M. C. PAYNE, M. P. TETER, D. C. ALLAN, T. A. ARIAS, AND J. D. JOANNOPOULOS, *Iterative minimization techniques for ab initio total-energy calculations: molecular dynamics and conjugate gradients*, Rev. Mod. Phys., 64 (1992), pp. 1045–1097.

[17] B. G. PFROMMER, J. DEMMEL, AND H. SIMON, *Unconstrained energy functionals for electronic structure calculations*, Journal of Computational Physics, 150 (1999), pp. 287 – 298.

[18] V. R. SAUNDERS AND I. H. HILLIER, *A levelshifting method for converging closed shell hartreefock wave functions*, International Journal of Quantum Chemistry, 7 (1973), pp. 699–705.

[19] W. SUN AND Y.-X. YUAN, *Optimization Theory and Methods*, vol. 1 of Springer Optimization and Its Applications, Springer, New York, 2006. Nonlinear programming.

[20] L. THOGERSEN, J. OLSEN, D. YEAGER, P. JORGENSEN, P. SALEK, AND T. HELGAKER, *The trust-region self-consistent field method: Towards a black-box optimization in hartree–fock and kohn–sham theories*, The Journal of Chemical Physics, 121 (2004), pp. 16–27.

[21] T. VAN VOORHIS AND M. HEAD-GORDON, *A geometric approach to direct minimization*, Molecular Physics, 100 (2002), pp. 1713–1721.

[22] J. VANDEVONDELE AND J. HUTTER, *An efficient orbital transformation method for electronic structure calculations*, The Journal of Chemical Physics, 118 (2003), pp. 4365–4369.

[23] Z. WEN AND W. YIN, *A feasible method for optimization with orthogonality constraints*, Mathematical Programming, (2012).

[24] 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 (2008/09), pp. 1773–1788.

[25] 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.

[26] C. Yang, J. C. Meza, and L.-W. Wang, *A constrained optimization algorithm for total energy minimization in electronic structure calculations*, J. Comput. Phys., 217 (2006), pp. 709–721.

[27] ———, *A trust region direct constrained minimization algorithm for the Kohn-Sham equation*, SIAM J. Sci. Comput., 29 (2007), pp. 1854–1875.