Mathematical Analysis and Numerical Approximations of Density Functional Theory Models for Metallic Systems

MATHEMATICAL ANALYSIS AND NUMERICAL APPROXIMATIONS OF DENSITY FUNCTIONAL THEORY MODELS FOR METALLIC SYSTEMS*

XIAOYING DAI†, STEFANO DE GIRONCOLI‡, BIN YANG§, AND AIHUI ZHOU†

Abstract. In this paper, we investigate the energy minimization model of the ensemble Kohn-Sham density functional theory for metallic systems, in which a pseudo-eigenvalue matrix and a general smearing approach are involved. We study the invariance and the existence of the minimizer of the energy functional. We propose an adaptive double step size strategy and the corresponding preconditioned conjugate gradient methods for solving the energy minimization model. Under some mild but reasonable assumptions, we prove the global convergence of our algorithms. Numerical experiments show that our algorithms are efficient, especially for large scale metallic systems. In particular, our algorithms produce convergent numerical approximations for some metallic systems, for which the traditional self-consistent field iterations fail to converge.

Key words. ensemble Kohn-Sham density functional theory, metallic systems, mathematical analysis, numerical approximation, precondtioned conjugate gradient method, convergence

AMS subject classifications. 65K10, 65N25, 49S05, 35P30

1. Introduction.

The Kohn-Sham density functional theory (DFT) is widely used in the electronic structure calculations [2, 4, 25, 30]. The underlying mathematical model is often formulated as either a nonlinear eigenvalue problem or an energy minimization problem with an unitary constraint. The most commonly used approach for computing the Kohn-Sham DFT model is to solve the nonlinear eigenvalue problem by using the self-consistent field (SCF) iterations. However, the convergence of the SCF iterations is not guaranteed and the performance of the SCF iterations is unpredictable, especially for large scale systems. Consequently, people turn to pay attention to investigating the constrained energy minimization problem (see, e.g., [9, 16, 35, 40, 41] and references therein).

We particularly note that the efficient numerical methods for the classical Kohn-Sham DFT model, in which occupation numbers are either 1 or 0, are inefficient or even invalid for metallic systems. The main reason is that the gap between the highest occupied state and the lowest unoccupied state for metallic systems is very small or absent. More precisely, the classical Kohn-Sham DFT model becomes ill-posed due to its difficulty to separate the occupied states and unoccupied states.

To provide a well-posed and efficient mathematical model for metallic systems, the unoccupied states have been incorporated into the classical Kohn-Sham DFT model and the fractional occupancies has been applied in computations. For instance, the

*This work was supported by the National Natural Science Foundation of China under grant 12021001, the National Key R & D Program of China under grants 2019YFA0709600 and 2019YFA0709601, and the CAS President's International Fellowship for Visiting Scientists under grants 2019VMA0029. de Gironcoli also acknowledges support from the European Union's Horizon 2020 research and innovation program (Grant No. 824143, MaX "MAterials design at the eXascale" Centre of Excellence).

†LSEC, Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China; and School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China. {daixy, azhou}@lsec.cc.ac.cn.

‡Scuola Internazionale Superiore di Studi Avanzati (SISSA) and CNR-IOM DEMOCRITOS Simulation Centre, Via Bononea 265, 34146 Trieste, Italy. degironc@sissa.it.

§NCMIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China. binyang@lsec.cc.ac.cn.

ensemble Kohn-Sham DFT (or the finite-temperature Kohn-Sham DFT) is developed (see, e.g., [22]), in which the associated total energy is a nonlinear functional of wavefunctions and pseudo-eigenvalues (or occupation numbers). We see that the ensemble Kohn-Sham DFT can be formulated as a nonlinear eigenvalue problem or a constrained energy minimization problem. It is not difficult to apply the SCF iteration approach for the classical Kohn-Sham DFT model to the ensemble Kohn-Sham DFT model. We understand that some preconditioners have been also constructed to accelerate the SCF iterations [19, 22, 24, 42]. Unfortunately, the convergence of the SCF iterations for the ensemble Kohn-Sham DFT is not guaranteed yet.

In the context of solving the constrained energy minimization problem of the ensemble Kohn-Sham DFT, different from the classical Kohn-Sham DFT, we need to treat the occupation numbers as additional variables. There are more challenges for designing and analyzing an efficient algorithm. For example, we observe that the unitary invariance of the energy functional is not clear and applying the unitary transformation to the Kohn-Sham orbitals may not produce the ground states. We also understand that it is necessary to calculate the Kohn-Sham orbitals exactly [22] and it is usually required to choose a good unitary transformation of the wavefunctions when designing an optimization algorithm. We refer to [17, 18, 22] for constructing the unitary transformation of the wavefunctions to make energy approximations decay. Ismail-Beigi et al. [21] suggested expressing the unitary transformation as $P = e^{\mathrm{i}B}$ and minimizing the energy functional with respect to the Hermitian matrix B. However, the unitary transformation is incorporated into the model when some matrix representations are applied. Marzari et al. [28] proposed an optimization algorithm by adopting a matrix representation of the occupation numbers, which we call the occupation matrix, and they got an unitarily invariant functional of wavefunctions by minimizing the occupation matrix. It is shown in [28] that it is not necessary to construct the unitary transformation. Later on, Freysoldt et al. [14] introduced the so-called pseudo-Hamitonian matrix and proposed a preconditioned conjuagte gradient (PCG) algorithm to minimize the energy functional with respect to the wavefunctions and the pseudo-Hamiltonian matrix, in which the unitary transformation is constructed automatically by minimizing the energy functional with respect to the pseudo-Hamiltonian matrix. Recently, Ulbrich et al. [37] studied a proximal gradient method for the ensemble Kohn-Sham DFT with the Fermi-Dirac smearing. We may refer to [1, 34] for more works on the direct minimization algorithms for the ensemble Kohn-Sham DFT model. To our knowledge, there is little mathematical analysis on the ensemble Kohn-Sham DFT and its approximations. In this paper, we investigate the energy minimization model of the ensemble Kohn-Sham DFT from a mathematical aspect, and design and analyze the associated optimization algorithms.

The rest of this paper is organized as follows. In the next section, we introduce some basic notation and the energy minimization model of the ensemble Kohn-Sham density functional theory with the pseudo-eigenvalue matrix and the general smearing method. In section 3, we study the invariance and the exsistence of the minimizer for the ensemble Kohn-Sham energy functional. In section 4, we propose an adaptive double step size strategy and the corresponding preconditioned conjugate gradient (PCG) algorithms to solve the energy minimization problem. Under some mild but reasonable assumptions, we then prove the global convergence of the PCG algorithms based on the adaptive double step size strategy we proposed. We report several numerical experiments in section 5 to demonstrate our theory and show the superiority of our algorithms over the traditional SCF iterations. We give some concluding remarks in section 6. Finally, we provide some details of the gradient of the energy

functional in Appendix A and the derivation process to get the standard Kohn-Sham equation in Appendix B.

2. Preliminaries.

2.1. Basic notation. Throughout this paper, we consider periodic systems. Since we usually apply a large enough unit cell when calculating isolated systems, our definitions and conclusions are applicable to the isolated systems in practice. Let $\Omega = \{x_1\xi_1 + x_2\xi_2 + x_3\xi_3 : x_1, x_2, x_3 \in [0,1)\}$ be the unit cell, where $\xi_1, \xi_2, \xi_3 \in \mathbb{R}^3$ are three non-coplanar vectors. Then the associated Bravais lattice and the reciprocal lattice are $\mathcal{R} = \{n_1\xi_1 + n_2\xi_2 + n_3\xi_3 : n_1, n_2, n_3 \in \mathbb{Z}\}$ and $\mathcal{R}^* = \{m_1\zeta_1 + m_2\zeta_2 + m_3\zeta_3 : m_1, m_2, m_3 \in \mathbb{Z}\}$, respectively. Here, $\mathbb{Z}$ represents the set of all integers and

$$\zeta_1 = 2\pi\frac{\xi_2 \times \xi_3}{\xi_1 \cdot (\xi_2 \times \xi_3)},\ \zeta_2 = 2\pi\frac{\xi_3 \times \xi_1}{\xi_2 \cdot (\xi_3 \times \xi_1)},\ \zeta_3 = 2\pi\frac{\xi_1 \times \xi_2}{\xi_3 \cdot (\xi_1 \times \xi_2)}.$$

For $\mathrm{G} \in \mathcal{R}^*$, we denote by $e_{\mathrm{G}}(r) = |\Omega|^{-1/2}e^{\mathrm{i}\mathrm{G}\cdot r}$ the planewave with wavevector G, where $|\Omega|$ is the volume of Ω. The family $\{e_{\mathrm{G}}\}_{G\in\mathcal{R}^*}$ forms an orthonormal basis of the complex valued $\mathcal{R}$-periodic functions space

$$L^2_{\#}(\Omega, \mathbb{C}) = \left\{\psi \in L^2_{\mathrm{loc}}(\mathbb{R}^3, \mathbb{C}) : \psi \text{ is } \mathcal{R}\text{-periodic}\right\}, \tag{2.1}$$

and for any $\psi \in L^2_{\#}(\Omega, \mathbb{C})$,

$$\psi(r) = \sum_{\mathrm{G}\in\mathcal{R}^*} \hat{\psi}_{\mathrm{G}} e_{\mathrm{G}}(r) \quad \text{with} \quad \hat{\psi}_{\mathrm{G}} = \frac{1}{|\Omega|^{\frac{1}{2}}}\int_{\Omega} \psi(r)e^{-\mathrm{i}\mathrm{G}\cdot r}\mathrm{d}r.$$

We define the Sobolev space of complex valued $\mathcal{R}$-periodic functions as

$$H^s_{\#}(\Omega, \mathbb{C}) = \left\{\psi \in L^2_{\#}(\Omega, \mathbb{C}) : \sum_{\mathrm{G}\in\mathcal{R}^*} (1 + |\mathrm{G}|^2)^s |\hat{\psi}_{\mathrm{G}}|^2 < \infty\right\}$$

with $s \in \mathbb{R}$, endowed with the inner product

$$(\psi, \phi)_{H^s_{\#}} = \sum_{\mathrm{G}\in\mathcal{R}^*} (1 + |\mathrm{G}|^2)^s \bar{\hat{\psi}}_{\mathrm{G}}\hat{\phi}_{\mathrm{G}},$$

and the induced norm

$$\|\psi\|^2_{H^s_{\#}} = \sum_{\mathrm{G}\in\mathcal{R}^*} (1 + |\mathrm{G}|^2)^s |\hat{\psi}_{\mathrm{G}}|^2.$$

For convenience, unless otherwise specified, $(\cdot,\cdot)$ and $\|\cdot\|$ always represent the inner product and the norm of $L^2_{\#}(\Omega, \mathbb{C})$, respectively.

Let $\Psi = (\psi_1, \ldots, \psi_N) \in (L^2_{\#}(\Omega, \mathbb{C}))^N, \Phi = (\phi_1, \ldots, \phi_N) \in (L^2_{\#}(\Omega, \mathbb{C}))^N$. Here N is some positive integer. We can view Ψ and Φ as vectors with elements being functions. Then we have

$$\Psi\Phi^* = \sum_{i=1}^{N} \psi_i\bar{\phi}_i,\ \Psi^*\Phi = (\bar{\psi}_i\phi_j)_{i,j=1}^N.$$

For any $A = (A_{ij})_{i,j=1}^N \in \mathbb{C}^{N\times N}$, we denote by

$$A\Psi^* = \left(\sum_{j=1}^{N} A_{1j}\bar{\psi}_j, \ldots, \sum_{j=1}^{N} A_{Nj}\bar{\psi}_j\right)^T,\ \Psi A = \left(\sum_{i=1}^{N} A_{i1}\psi_i, \ldots, \sum_{i=1}^{N} A_{iN}\psi_i\right).$$

Define

$$\langle \Psi^* \Phi \rangle = ((\psi_i, \phi_j))_{i,j=1}^N \in \mathbb{C}^{N\times N}.$$

For any positive integer n and any $\Psi = (\Psi_1, \Psi_2, \ldots, \Psi_n),\ \Phi = (\Phi_1, \Phi_2, \ldots, \Phi_n) \in ((L^2(\Omega, \mathbb{C}))^N)^n$, we define its inner product as $\langle \Psi, \Phi \rangle = \sum_{i=1}^n \mathrm{tr}\langle \Psi_i^* \Phi_i \rangle$. The induced norm is $\|\Psi\| = \sqrt{\langle \Psi, \Psi \rangle}$. We shall use the notation

$$\|\Psi\|_\infty = \max_{i=1,2,\ldots,n} \|\Psi_i\|$$

for convenience.

For any $A = (A_1, A_2, \ldots, A_n), B = (B_1, B_2, \ldots, B_n) \in (\mathbb{C}^{N\times N})^n$, we define its inner product as $\langle A, B \rangle = \sum_{i=1}^n \mathrm{tr}(A_i^* B_i)$. And the induced norm is Frobenius norm, denoted by $\|\cdot\|_F$. We shall use the notation $\|\cdot\|_{sF}$ defined as

$$\|A\|_{sF} = \min_{c\in\mathbb{C}} \|cI_N - A\|_F,$$

where $cI_N - A \coloneqq (cI_N - A_1, cI_N - A_2, \ldots, cI_N - A_n)$. It is easy to obtain

$$\|A\|_{sF} = \left\| \frac{\sum_{i=1}^n \mathrm{tr}\, A_i}{nN} I_N - A \right\|_F. \tag{2.2}$$

Define

$$\|A\|_{sF,\infty} = \min_{i=1,2,\ldots,n} \|c_A I_N - A_i\|_F,$$

where $c_A = \frac{\sum_{i=1}^n \mathrm{tr}\, A_i}{nN}$. It is easy to get the following properties for $\|\cdot\|_{sF}$ by (2.2).

PROPOSITION 2.1. *Let $A, B \in (\mathbb{C}^{N\times N})^n$, then the following properties of $\|\cdot\|_{sF}$ hold true:*

1. $\|A - B\|_{sF} = 0$ *if and only if there exists $c \in \mathbb{C}$ such that $A = B + cI_N$;*
2. $\|\cdot\|_{sF}$ *satisfies the triangle inequality, i.e., $\|A + B\|_{sF} \le \|A\|_{sF} + \|B\|_{sF}$;*
3. $\|\cdot\|_{sF}$ *satisfies the absolute homogeneity, i.e., $\|\alpha A\|_{sF} = |\alpha| \|A\|_{sF}$ for any $\alpha \in \mathbb{C}$;*
4. *if* $\sum_{i=1}^n \mathrm{tr}\, A_i = 0$, *then*

$$|\langle A, B \rangle| \le \|A\|_{sF} \|B\|_{sF}.$$

It follows from Proposition 2.1 and (2.2) that $\|A\|_{sF}$ is the norm of the following linear space

$$\left\{ A = (A_1, A_2, \ldots, A_n) \in (\mathbb{C}^{N\times N})^n : \sum_{i=1}^n \mathrm{tr}\, A_i = 0 \right\}.$$

The Stiefel manifold is defined by

$$\mathcal{M}_{\mathcal{B},\mathbb{C}}^N = \{\Psi \in (H_\#^1(\Omega, \mathbb{C}))^N : \langle \Psi^* \mathcal{B} \Psi \rangle = I_N\},$$

where $\mathcal{B}$: $(L^2(\Omega, \mathbb{C}))^N \to (L^2(\Omega, \mathbb{C}))^N$ is a bounded and self-adjoint operator. Let

$$\mathcal{O}_{\mathbb{C}}^{N\times N} = \{P \in \mathbb{C}^{N\times N} : P^* P = I_N\},\ \mathcal{S}_{\mathbb{C}}^{N\times N} = \{A \in \mathbb{C}^{N\times N} : A^* = A\}.$$

If only real values are taken into account, we then remove $\mathbb{C}$ or replace $\mathbb{C}$ with $\mathbb{R}$ and replace the conjugate transpose symbol $*$ by the transpose symbol T in the above

notation. We note that the Fourier coefficients of real valued $\mathcal{R}$-periodic functions have some symmetry, more precisely,

$$H^s_{\#}(\Omega) = \left\{\psi \in H^s_{\#}(\Omega,\mathbb{C}) : \forall \mathrm{G} \in \mathcal{R}^*,\ \hat{\psi}_{-\mathrm{G}} = \bar{\hat{\psi}}_{\mathrm{G}}\right\}. \tag{2.3}$$

We then introduce some projections of wavefunctions. Let $\Psi \in \mathcal{M}^N_{\mathcal{B},\mathbb{C}}$. We know that the tangent space of $\mathcal{M}^N_{\mathcal{B},\mathbb{C}}$ at Ψ is

$$\mathcal{T}_\Psi \mathcal{M}^N_{\mathcal{B}} = \{\Phi \in (H^1_{\#}(\Omega,\mathbb{C}))^N : \langle \Phi^* \mathcal{B}\Psi\rangle + \langle \Psi^* \mathcal{B}\Phi\rangle = 0 \in \mathbb{C}^{N\times N}\}.$$

Let

$$K_\Psi = \{\Phi \in (H^1_{\#}(\Omega,\mathbb{C}))^N : \langle \Phi^* \Psi\rangle + \langle \Psi^* \Phi\rangle = 0 \in \mathbb{C}^{N\times N}\}.$$

It is clear that $\mathcal{T}_\Psi \mathcal{M}^N_{\mathcal{B},\mathbb{C}} = K_\Psi$ provided $\mathcal{B} = \mathcal{I}$, where $\mathcal{I}$ is the identity operator. For any $\alpha \in \mathbb{R}$, we define the linear operator onto K_Ψ by

$$P_{\alpha,\Psi}(\Phi) = (\Phi - \mathcal{B}\Psi\langle\Psi^*\Phi\rangle) + \alpha\mathcal{B}\Psi(\langle\Psi^*\Phi\rangle - \langle\Phi^*\Psi\rangle),\ \forall \Phi \in (H^1_{\#}(\Omega,\mathbb{C}))^N. \tag{2.4}$$

We see that

$$P^2_{\alpha,\Psi}(\Phi) = P_{\alpha,\Psi}(\Phi) + \alpha(2\alpha-1)\mathcal{B}\Psi(\langle\Psi^*\Phi\rangle - \langle\Phi^*\Psi\rangle),\ \forall \Phi \in (H^1_{\#}(\Omega,\mathbb{C}))^N,$$

which indicates that $P_{\alpha,\Psi}$ is a projection if and only if $\alpha = 0$ or $1/2$. Define

$$P^*_{\alpha,\Psi}(\Phi) = (\Phi - \Psi\langle\Psi^*\mathcal{B}\Phi\rangle) + \alpha\Psi(\langle\Psi^*\mathcal{B}\Phi\rangle - \langle\Phi^*\mathcal{B}\Psi\rangle),\ \forall \Phi \in (H^1_{\#}(\Omega,\mathbb{C}))^N.$$

We have that for any $\Phi_1, \Phi_2 \in (H^1_{\#}(\Omega,\mathbb{C}))^N$, $\langle P_{\alpha,\Psi}(\Phi_1), \Phi_2\rangle$ and $\left\langle \Phi_1, P^*_{\alpha,\Psi}(\Phi_2)\right\rangle$ have the same real part. Thus $P^*_{\alpha,\Psi}$ is the adjoint operator of $P_{\alpha,\Psi}$ if only real functions are involved. We mention that $P_{0,\Psi}(\Phi)$ is orthogonal to Ψ for any $\Phi \in (H^1_{\#}(\Omega,\mathbb{C}))^N$.

2.2. Ensemble Kohn-Sham DFT model for metallic systems. We consider the ensemble Kohn-Sham density functional theory, in which we adopt the matrix representation of occupations [14, 28]. We see from Bloch's theorem [25] that the kinetic energy and the electronic density are given by the integral over the Brillouin zone (BZ). If BZ sampling is used to discrete the integral over BZ, the ensemble Kohn-Sham energy functional with a general smearing approach can be formulated as

$$\mathcal{F}(\Psi,\eta) = \mathcal{E}(\Psi,\eta) - \sigma\sum_{k\in\mathcal{K}} w_{\mathrm{k}}\,\mathrm{tr}\,S\left(\frac{1}{\sigma}(\eta_{\mathrm{k}} - \mu I_N)\right) \tag{2.5}$$

with wavefunctions $\Psi = (\Psi_{\mathrm{k}})_{\mathrm{k}\in\mathcal{K}} \in ((H^1_{\#}(\mathbb{R}^3,\mathbb{C}))^N)^{|\mathcal{K}|}$ and the pseudo-eigenvalue matrices $\eta = (\eta_{\mathrm{k}})_{\mathrm{k}\in\mathcal{K}} \in (\mathcal{S}^{N\times N}_{\mathbb{C}})^{|\mathcal{K}|}$, where

$$\begin{aligned}\mathcal{E}(\Psi,\eta) = &\sum_{\mathrm{k}\in\mathcal{K}} w_{\mathrm{k}}\,\mathrm{tr}\left(\left\langle \Psi^*_{\mathrm{k}}\left(-\frac{1}{2}(\mathrm{i}\mathrm{k}+\nabla)^2 + V_{\mathrm{nl}}\right)\Psi_{\mathrm{k}}\right\rangle F_{\eta_{\mathrm{k}}}\right) + \int_\Omega V_{\mathrm{loc}}(r)\rho_{\Psi,\eta}(r)\mathrm{d}r\\ &+ \frac{1}{2}\int_\Omega\int_\Omega \frac{\rho_{\Psi,\eta}(r)\rho_{\Psi,\eta}(r')}{|r-r'|}\mathrm{d}r\mathrm{d}r' + \mathcal{E}_{\mathrm{xc}}(\rho_{\Psi,\eta}).\end{aligned}$$

Here $\mathcal{K}$ is a finite subset of BZ, w_{k} is the weight associated to k-points $\mathrm{k} \in \mathcal{K}$ satisfying

$$\sum_{\mathrm{k}\in\mathcal{K}} w_{\mathrm{k}} = 2,$$

N is the number of wavefunctions for one k-point, $\sigma = k_B T$ with the Boltzmann constant k_B and the temperature T,

$$F_{\eta_{\mathrm{k}}} = f\left(\frac{1}{\sigma}(\eta_{\mathrm{k}} - \mu I_N)\right),$$

f is a function which is sometimes called the smearing function, and μ is a function of η which will be determined later, S is a function associated to the entropy term. The electronic density $\rho_{\Psi,\eta}$ is

$$\rho_{\Psi,\eta} = \sum_{\mathrm{k}\in\mathcal{K}} w_{\mathrm{k}} \operatorname{tr}((\Psi_{\mathrm{k}}^* \Psi_{\mathrm{k}} + \langle \Psi_{\mathrm{k}}^* M\rangle \mathcal{Q} \langle M^* \Psi_{\mathrm{k}}\rangle) F_{\eta_{\mathrm{k}}})$$

with $M = (\varphi_1, \ldots, \varphi_K) \in (L^2_{\#}(\Omega, \mathbb{C}))^K$ and the Hermitian-matrix-valued function $\mathcal{Q} = (\mathcal{Q}_{ij})_{i,j=1}^K \in (L^2_{\#}(\Omega, \mathbb{C}))^{K\times K}$. Sometimes we shall simply denote $\rho_{\Psi,\eta}$ by ρ. $V_{\mathrm{loc}} \in L^2_{\#}(\Omega, \mathbb{C})$ is the local pseudopotential and V_{nl} is the nonlocal pseudopotential defined by $\Psi_{\mathrm{k}} \mapsto V_{\mathrm{nl}}(\Psi_{\mathrm{k}}) = MD\langle M^* \Psi_{\mathrm{k}}\rangle$ with $D \in \mathcal{S}_{\mathbb{C}}^{K\times K}$. Note that the form of (2.5) is suitable for the full potential calculations, the pseudopotential approximations [36, 38] and the projector augmented wave (PAW) method [3]. For instance, if the norm-conserving pseudopotential is applied, then $\mathcal{Q} = 0$ and $\rho_{\Psi,\eta} = \sum_{\mathrm{k}\in\mathcal{K}} w_{\mathrm{k}} \operatorname{tr}(\Psi_{\mathrm{k}}^* \Psi_{\mathrm{k}})$. In theory, N should be $+\infty$ for the ensemble Kohn-Sham DFT. However, N has to be set to be finite in practice. We require $N > N_b$ where N_b is the number or the half number of electrons. For example, in Quantum ESPRESSO, N is set to $N_b + \lfloor 0.2 N_b \rfloor$ by default, where $\lfloor x \rfloor$ is the greatest integer not larger than x.

Now we address the function μ of η in detail. Assume that f and S satisfy the following properties:

A.I f and S are analytic functions on $\mathbb{R}$ satisfying $S'(x) = x f'(x)$.

A.II $\lim_{x\to-\infty} f(x) = 1$ and $\lim_{x\to+\infty} f(x) = 0$.

A.III $\lim_{x\to+\infty} S(x)$ and $\lim_{x\to-\infty} S(x)$ exist.

A.IV f is strictly monotonically decreasing.

Under these assumptions, for given $\eta \in \left(\mathcal{S}_{\mathbb{C}}^{N\times N}\right)^{|\mathcal{K}|}$, there is one and only one $\mu \in \mathbb{R}$ satisfying $\sum_{\mathrm{k}\in\mathcal{K}} w_{\mathrm{k}} \operatorname{tr} F_{\eta_{\mathrm{k}}} = N_e$. Here N_e is the number of electrons. Thus, we choose μ in (2.5) as the unique function of η from $\left(\mathcal{S}_{\mathbb{C}}^{N\times N}\right)^{|\mathcal{K}|}$ to $\mathbb{R}$ such that $\sum_{\mathrm{k}\in\mathcal{K}} w_{\mathrm{k}} \operatorname{tr} F_{\eta_{\mathrm{k}}} = N_e$.

We list several possible choices for the smearing function used in the literature.

- the Fermi-Dirac smearing [5]:

$$f_{\mathrm{FD}}(x) = \frac{1}{1+e^x},\ S_{\mathrm{FD}}(x) = -[f_{\mathrm{FD}}(x)\ln f_{\mathrm{FD}}(x) + (1-f_{\mathrm{FD}}(x))\ln(1-f_{\mathrm{FD}}(x))].$$

- the Gaussian smearing [12, 15]:

$$f_{\mathrm{GS}}(x) = \frac{1}{2}(1 - \operatorname{erf}(x)),\ S_{\mathrm{GS}}(x) = \frac{1}{2\sqrt{\pi}} e^{-x^2}.$$

- the Methfessel-Paxton smearing [29]:

$$f_{\mathrm{MP},m}(x) = f_{\mathrm{GS}}(x) + \sum_{i=1}^{m} A_i H_{2i-1}(x) e^{-x^2},\ S_{\mathrm{MP},m}(x) = \frac{1}{2} A_m H_{2m}(x) e^{-x^2},$$

where H_i are the Hermite polynomials (defined as $H_0(x) = 1$, $H_{i+1}(x) = 2xH_i(x) - H_i'(x)$) and

$$A_i = \frac{(-1)^i}{i!4^i\sqrt{\pi}}.$$

- the Marzari-Vanderbilt smearing [26, 27]:

$$f_{\mathrm{MV}}(x) = f_{\mathrm{GS}}(x) + \frac{1}{4\sqrt{\pi}}\left(-\frac{1}{2}aH_2(x) + H_1(x)\right)e^{-x^2},$$

$$S_{\mathrm{MV}}(x) = \frac{1}{4\sqrt{\pi}}\left(-\frac{1}{2}H_2(x) + ax^2H_1(x)\right)e^{-x^2},$$

where a is a free parameter such that $f_{\mathrm{MV}}(x)$ is nonnegative for any $x \in \mathbb{R}$. Marzari suggests choosing $a = -0.5634$ or $a = -\sqrt{2/3}$ in [26].

We see that the assumptions A.I-A.II imply the existence of $\mu \in \mathbb{R}$ such that $\sum_{\mathrm{k}\in\mathcal{K}} w_{\mathrm{k}} \operatorname{tr} F_{\eta_{\mathrm{k}}} = N_e$ for any given $\eta \in \left(\mathcal{S}_{\mathbb{C}}^{N\times N}\right)^{|\mathcal{K}|}$. Further, if A.IV is satisfied, then μ is unique. Thus, μ is a function of η when the Fermi-Dirac smearing and the Gaussian smearing are applied. But for some other smearing such as the Methfessel-Paxton smearing and the Marzari-Vanderbilt smearing, it is still open whether μ is unique. In practice, we will always assume that μ is a function of η such that $\sum_{\mathrm{k}\in\mathcal{K}} w_{\mathrm{k}} \operatorname{tr} F_{\eta_{\mathrm{k}}} = N_e$.

According to the ensemble Kohn-Sham DFT, we solve the following constrained minimization problem

$$\inf_{(\Psi,\eta)\in\left(\mathcal{M}_{\mathcal{B},\mathbb{C}}^{N}\right)^{|\mathcal{K}|}\times\left(\mathcal{S}_{\mathbb{C}}^{N\times N}\right)^{|\mathcal{K}|}} \mathcal{F}(\Psi,\eta) \tag{2.6}$$

to obtain the ground state of the system, where $\mathcal{B}$ is an operator defined by $\Psi \mapsto \mathcal{B}\Psi = \Psi + MQ\langle M^*\Psi\rangle$ with $Q = \int_\Omega \mathcal{Q}(r)\mathrm{d}r$. Note that $\mathcal{B}$ is bounded and self-adjoint. The associated Lagrange functional is

$$\mathcal{L}(\Psi,\eta,\Lambda) = \mathcal{F}(\Psi,\eta) - \sum_{\mathrm{k}\in\mathcal{K}} w_{\mathrm{k}} \operatorname{tr}[\Lambda_{\mathrm{k}}^*(\langle\Psi_{\mathrm{k}}^*\mathcal{B}\Psi_{\mathrm{k}}\rangle - I_N)] \tag{2.7}$$

with the Lagrange multiplier $\Lambda = (\Lambda_{\mathrm{k}})_{\mathrm{k}\in\mathcal{K}} \in \left(\mathbb{C}^{N\times N}\right)^{|\mathcal{K}|}$. Note that throughout this paper, since our discussion with respect to η is in the linear space $\left(\mathcal{S}_{\mathbb{C}}^{N\times N}\right)^{|\mathcal{K}|}$ over $\mathbb{R}$, there is no term associated with the constraint $\eta \in \left(\mathcal{S}_{\mathbb{C}}^{N\times N}\right)^{|\mathcal{K}|}$ in the Lagrange functional (2.7).

Assume that the exchange-correction functional $\mathcal{E}_{\mathrm{xc}}$ is differentiable. We regard Ψ_{k} and $\bar{\Psi}_{\mathrm{k}}$ as two independent variables for all $\mathrm{k} \in \mathcal{K}$ and view $\mathcal{F}$ as a functional of Ψ, $\bar{\Psi}$ and η. Then we get (see Appendix A)

$$\mathcal{F}_{\Psi_{\mathrm{k}}}(\Psi,\eta) = w_{\mathrm{k}}H_{\mathrm{k}}(\rho_{\Psi,\eta})\Psi_{\mathrm{k}}F_{\eta_{\mathrm{k}}}$$

and

$$\mathcal{L}_{\Psi_{\mathrm{k}}}(\Psi,\eta,\Lambda) = w_{\mathrm{k}}(H_{\mathrm{k}}(\rho_{\Psi,\eta})\Psi_{\mathrm{k}}F_{\eta_{\mathrm{k}}} - \mathcal{B}\Psi_{\mathrm{k}}\Lambda_{\mathrm{k}}), \tag{2.8}$$

where $\mathcal{F}_{\Psi_{\mathrm{k}}}$ and $\mathcal{L}_{\Psi_{\mathrm{k}}}$ are Wirtinger derivatives,

$$H_{\mathrm{k}}(\rho) = -\frac{1}{2}(\mathrm{i}\mathrm{k} + \nabla)^2 + \tilde{V}_{\mathrm{loc}}(\rho) + \tilde{V}_{\mathrm{nl}}(\rho)$$

with $\tilde{V}_{\rm loc}(\rho) = V_{\rm loc} + \displaystyle\int_\Omega \frac{\rho(r)}{|\cdot - r|}{\rm d}r + V_{\rm xc}(\rho)$, $\tilde{V}_{\rm nl}(\rho) : \Psi_{\rm k} \mapsto V_{\rm nl}(\Psi_{\rm k}) + M\tilde{D}\,\langle M^*\Psi_{\rm k}\rangle$, $V_{\rm xc}(\rho) = \dfrac{\delta \mathcal{E}_{\rm xc}}{\delta \rho}$, and

$$\tilde{D} = \int_\Omega \tilde{V}_{\rm loc}(\rho)(r)\mathcal{Q}(r)\,{\rm d}r \in \mathcal{S}_{\mathbb{C}}^{K\times K}.$$

Here we use the convenient notation $\mathcal{F}(\Psi,\eta) = \mathcal{F}(\Psi,\bar{\Psi},\eta)$ and $\mathcal{L}(\Psi,\eta,\Lambda) = \mathcal{L}(\Psi,\bar{\Psi},\eta,\Lambda)$. Set

$$\nabla_{\Psi_{\rm k}}\mathcal{F}(\Psi,\eta) = 2w_{\rm k}(H_{\rm k}(\rho_{\Psi,\eta})\Psi_{\rm k} - \mathcal{B}\Psi_{\rm k}\langle\Psi_{\rm k}^* H(\rho_{\Psi,\eta})\Psi_{\rm k}\rangle)F_{\eta_{\rm k}}$$

and $\nabla_\Psi\mathcal{F} = (\nabla_{\Psi_{\rm k}}\mathcal{F})_{{\rm k}\in\mathcal{K}}$. Given η, we denote by $\nabla_{\eta_{\rm k}}\mathcal{F} = \mathcal{F}_{\eta_{\rm k}}^T$ and $\nabla_\eta\mathcal{F} = (\nabla_{\eta_{\rm k}}\mathcal{F})_{{\rm k}\in\mathcal{K}}$, where

$$\mathcal{F}_{\eta_{\rm k}} = \left(\frac{\partial\mathcal{F}}{\partial\eta_{{\rm k}ij}}\right)_{i,j=1}^N.$$

When all $\eta_{\rm k}$ are diagonal matrices, $\dfrac{\partial\mathcal{F}}{\partial\eta_{{\rm k}ij}}$ is given by

$$\begin{aligned}\frac{\partial\mathcal{F}}{\partial\eta_{{\rm k}ij}} = w_{\rm k}\Bigg(&(\langle\psi_{{\rm k}i}, H_{\rm k}(\rho_{\Psi,\eta})\psi_{{\rm k}i}\rangle - \epsilon_{{\rm k}i})\frac{1}{\sigma}f'\left(\frac{\epsilon_{{\rm k}i}-\mu}{\sigma}\right)\delta_{ij}\\ &- \frac{f'\left(\frac{\epsilon_{{\rm k}'i}-\mu}{\sigma}\right)\delta_{ij}}{\sum_{{\rm k}'} w_{{\rm k}'}\sum_{i'=1}^N f'\left(\frac{\epsilon_{{\rm k}'i'}-\mu}{\sigma}\right)}d_\mu\\ &+ \langle\psi_{{\rm k}j}, H(\rho_{\Psi,\eta})\psi_{{\rm k}i}\rangle\frac{f_{{\rm k}j}-f_{{\rm k}i}}{\epsilon_{{\rm k}j}-\epsilon_{{\rm k}i}}(1-\delta_{ij})\Bigg)\end{aligned}$$

for any ${\rm k}\in\mathcal{K}$, where $\Psi_{\rm k} = (\psi_{{\rm k}1},\psi_{{\rm k}2},\ldots,\psi_{{\rm k}N})$, $\eta_{\rm k} = {\rm Diag}(\epsilon_{{\rm k}1},\epsilon_{{\rm k}2},\ldots,\epsilon_{{\rm k}N})$, $f_{{\rm k}i} = f((\epsilon_{{\rm k}i}-\mu)/\sigma)$, $\frac{f_{{\rm k}j}-f_{{\rm k}i}}{\epsilon_{{\rm k}j}-\epsilon_{{\rm k}i}} = \frac{1}{\sigma}f\left(\frac{\epsilon_{{\rm k}i}-\mu}{\sigma}\right)$ provided $\epsilon_{{\rm k}j} = \epsilon_{{\rm k}i}$,

$$d_\mu = \sum_{{\rm k}'\in\mathcal{K}} w_{{\rm k}'}\sum_{i'=1}^N(\langle\psi_{{\rm k}'i'}, H_{{\rm k}'}(\rho_{\Psi,\eta})\psi_{{\rm k}'i'}\rangle - \epsilon_{{\rm k}'i'})\frac{1}{\sigma}f'\left(\frac{\epsilon_{{\rm k}'i'}-\mu}{\sigma}\right).$$

It is clear that $\mathcal{L}_{\Psi_{\rm k}}(\Psi,\eta,\Lambda) = 0$ and $\mathcal{L}_{\eta_{\rm k}}(\Psi,\eta,\Lambda) = 0$ for all ${\rm k}\in\mathcal{K}$ mean that $\nabla_\Psi\mathcal{F}(\Psi,\eta) = 0$ and $\nabla_\eta\mathcal{F}(\Psi,\eta) = 0$. And $\nabla_\Psi\mathcal{F}(\Psi,\eta) = 0$ and $\nabla_\eta\mathcal{F}(\Psi,\eta) = 0$ mean that there exists some Λ such that $\mathcal{L}_{\Psi_{\rm k}}(\Psi,\eta,\Lambda) = 0$ and $\mathcal{L}_{\eta_{\rm k}}(\Psi,\eta,\Lambda) = 0$ for all ${\rm k}\in\mathcal{K}$. As for the classical Kohn-Sham DFT model, let $\mathcal{L}_\Psi(\Phi,\eta,\Lambda) = 0$ and $\mathcal{L}_\eta(\Phi,\eta,\Lambda) = 0$, we will obtain the standard Kohn-Sham equation (see Appendix B for details).

3. Mathematical analysis. In this section, we investigate some basic mathematical properties of the ensemble Kohn-Sham DFT model, including the invariance and the existence of the minimizer of the energy functional.

3.1. Invariance. We first have the following invariance of the energy functional.

THEOREM 3.1. *For any $c\in\mathbb{R}$, $(\Psi,\eta) := (\Psi_{\rm k},\eta_{\rm k})_{{\rm k}\in\mathcal{K}} \in ((H^1_\#(\Omega,\mathbb{C}))^N)^{|\mathcal{K}|}\times\left(\mathcal{S}_{\mathbb{C}}^{N\times N}\right)^{|\mathcal{K}|}$ and $P := (P_{\rm k})_{{\rm k}\in\mathcal{K}} \in \left(\mathcal{O}_{\mathbb{C}}^{N\times N}\right)^{|\mathcal{K}|}$, there holds*

$$\mathcal{F}(\Psi P, P^*(\eta + cI_N)P) = \mathcal{F}(\Psi,\eta), \tag{3.1}$$

where $\Psi P = (\Psi_{\rm k}P_{\rm k})_{{\rm k}\in\mathcal{K}}$, $P^*\eta P = (P_{\rm k}^*\eta_{\rm k}P_{\rm k})_{{\rm k}\in\mathcal{K}}$.

Proof. It is sufficient to prove that

$$\mathcal{F}(\Psi, \eta + cI_N) = \mathcal{F}(\Psi, \eta), \tag{3.2}$$

$$\mathcal{F}(\Psi P, P^*\eta P) = \mathcal{F}(\Psi, \eta) \tag{3.3}$$

hold true for any $c \in \mathbb{R}$, $(\Psi, \eta) \in ((H^1_{\#}(\Omega, \mathbb{C}))^N)^{|\mathcal{K}|} \times \left(\mathcal{S}^{N\times N}_{\mathbb{C}}\right)^{|\mathcal{K}|}$ and $P \in \left(\mathcal{O}^{N\times N}_{\mathbb{C}}\right)^{|\mathcal{K}|}$.

We first prove the equation (3.2). By the uniqueness of μ that $\sum_{\mathrm{k}\in\mathcal{K}} w_{\mathrm{k}} \operatorname{tr} F_{\eta_{\mathrm{k}}} = N_e$, we obtain $\mu(\eta+cI_N) = \mu(\eta)+c$ for any $c \in \mathbb{R}$. Thus, we have $(F_{\eta_{\mathrm{k}}+cI_N})_{\mathrm{k}\in\mathcal{K}} = (F_{\eta_{\mathrm{k}}})_{\mathrm{k}\in\mathcal{K}}$ and

$$\left(S\left(\frac{1}{\sigma}(\eta_{\mathrm{k}} + cI_N - \mu(\eta + cI_N)I_N)\right)\right)_{\mathrm{k}\in\mathcal{K}} = \left(S\left(\frac{1}{\sigma}(\eta_{\mathrm{k}} - \mu(\eta)I_N)\right)\right)_{\mathrm{k}\in\mathcal{K}},$$

which lead to $\rho_{\Psi,\eta+cI_N} = \rho_{\Psi,\eta}$ and arrive at (3.2).

Next we prove the equation (3.3). Since f and S are analytic on $\mathbb{R}$, we have

$$P_{\mathrm{k}} f(\eta_{\mathrm{k}}) P_{\mathrm{k}}^* = f(P_{\mathrm{k}}\eta_{\mathrm{k}}P_{\mathrm{k}}^*),\ P_{\mathrm{k}} S(\eta_{\mathrm{k}}) P_{\mathrm{k}}^* = S(P_{\mathrm{k}}\eta_{\mathrm{k}}P_{\mathrm{k}}^*).$$

By the uniqueness of μ that $\sum_{\mathrm{k}\in\mathcal{K}} w_{\mathrm{k}} \operatorname{tr} F_{\eta_{\mathrm{k}}} = N_e$, we get $\mu(P^*\eta P) = \mu(\eta)$ for any $P \in (\mathcal{O}^{N\times N}_{\mathbb{C}})^{\mathcal{K}}$. Note that

$$\begin{aligned}
\rho_{\Psi P,\eta} &= \sum_{\mathrm{k}\in\mathcal{K}} w_{\mathrm{k}} \operatorname{tr}(P_{\mathrm{k}}^*(\Psi_{\mathrm{k}}^*\Psi_{\mathrm{k}} + \langle\Psi_{\mathrm{k}}^* M\rangle \mathcal{Q} \langle M^*\Psi_{\mathrm{k}}\rangle) P_{\mathrm{k}} F_{\eta_{\mathrm{k}}}) \\
&= \sum_{\mathrm{k}\in\mathcal{K}} w_{\mathrm{k}} \operatorname{tr}((\Psi_{\mathrm{k}}^*\Psi_{\mathrm{k}} + \langle\Psi_{\mathrm{k}}^* M\rangle \mathcal{Q} \langle M^*\Psi_{\mathrm{k}}\rangle) F_{P_{\mathrm{k}}\eta_{\mathrm{k}}P_{\mathrm{k}}^*}) \\
&= \rho_{\Psi, P\eta P^*}.
\end{aligned}$$

We have

$$\begin{aligned}
\mathcal{F}(\Psi P, \eta) &= \sum_{\mathrm{k}\in\mathcal{K}} w_{\mathrm{k}} \operatorname{tr}\left(\left\langle (\Psi_{\mathrm{k}}P_{\mathrm{k}})^* \left(-\frac{1}{2}\Delta + V_{\mathrm{nl}}\right)(\Psi_{\mathrm{k}}P_{\mathrm{k}})\right\rangle F_{\eta_{\mathrm{k}}}\right) \\
&\quad + \int_\Omega V_{\mathrm{loc}}(r)\rho_{\Psi P,\eta}(r)dr + \mathcal{E}_{\mathrm{HXC}}(\rho_{\Psi P,\eta}) - \sigma \sum_{\mathrm{k}\in\mathcal{K}} w_{\mathrm{k}} \operatorname{tr} P_{\mathrm{k}} S\left(\frac{1}{\sigma}(\eta_{\mathrm{k}} - \mu I)\right) P_{\mathrm{k}}^* \\
&= \sum_{\mathrm{k}\in\mathcal{K}} w_{\mathrm{k}} \operatorname{tr}\left(\left\langle \Psi_{\mathrm{k}}^* \left(-\frac{1}{2}(i\mathrm{k} + \nabla)^2 + V_{\mathrm{nl}}\right)\Psi_{\mathrm{k}}\right\rangle F_{P_{\mathrm{k}}\eta_{\mathrm{k}}P_{\mathrm{k}}^*}\right) \\
&\quad + \int_{\mathbb{R}^3} V_{\mathrm{loc}}(r)\rho_{\Psi,P\eta P^*}(r)dr + \mathcal{E}_{\mathrm{HXC}}(\rho_{\Psi,P\eta P^*}) - \sigma \sum_{\mathrm{k}\in\mathcal{K}} w_{\mathrm{k}} \operatorname{tr} S\left(\frac{1}{\sigma}(P_{\mathrm{k}}\eta_{\mathrm{k}}P_{\mathrm{k}}^* - \mu I)\right),
\end{aligned}$$

where

$$\mathcal{E}_{\mathrm{HXC}}(\rho_{\Psi,\eta}) = \frac{1}{2}\int_\Omega\int_\Omega \frac{\rho_{\Psi,\eta}(r)\rho_{\Psi,\eta}(r')}{|r-r'|}\mathrm{d}r\mathrm{d}r' + \mathcal{E}_{\mathrm{xc}}(\rho_{\Psi,\eta}),$$

namely,

$$\mathcal{F}(\Psi P, \eta) = \mathcal{F}(\Psi, P\eta P^*). \tag{3.4}$$

Finally we obtain from (3.4) that

$$\mathcal{F}(\Psi P, P^*\eta P) = \mathcal{F}(\Psi, P(P^*\eta P)P^*) = \mathcal{F}(\Psi, \eta). \qquad \square$$

We may view (3.2) as the translation invariance and (3.3) as the quasi unitary invariance.

We obtain from (3.1) that

$$\inf_{(\Psi,\eta)\in\left(\mathcal{M}^N_{\mathcal{B},\mathbb{C}}\right)^{|\mathcal{K}|}\times\left(\mathcal{D}^{N\times N}\right)^{|\mathcal{K}|}}\mathcal{F}(\Psi,\eta)=\inf_{(\Psi,\eta)\in\left(\mathcal{M}^N_{\mathcal{B},\mathbb{C}}\right)^{|\mathcal{K}|}\times\left(\mathcal{S}^{N\times N}_{\mathbb{C}}\right)^{|\mathcal{K}|}}\mathcal{F}(\Psi,\eta),\tag{3.5}$$

where $\mathcal{D}^{N\times N}=\{A\in\mathbb{R}^{N\times N}: A \text{ is a diagonal matrix}\}$. We see that

$$\inf_{(\Psi,\eta)\in\left(\mathcal{M}^N_{\mathcal{B},\mathbb{C}}\right)^{|\mathcal{K}|}\times\left(\mathcal{D}^{N\times N}\right)^{|\mathcal{K}|}}\mathcal{F}(\Psi,\eta)$$

is the original ensemble Kohn-Sham DFT model, which means that the model (2.6) is equivalent to the original ensemble Kohn-Sham DFT model.

We see from (3.1) that the solution of (2.6) is not unique. Thus we may turn to consider the following optimization problem

$$\inf_{[\Psi,\eta]\in\left(\mathcal{M}^N_{\mathcal{B},\mathbb{C}}\right)^{|\mathcal{K}|}\times\left(\mathcal{S}^{N\times N}_{\mathbb{C}}\right)^{|\mathcal{K}|}/\sim}\mathcal{F}(\Psi,\eta)\tag{3.6}$$

which is equivalent to (2.6). Here $\sim$ denotes the equivalence relation defined as follows: $(\Psi,\eta)\sim(\Psi',\eta')$ if and only if there exist $P\in\left(\mathcal{O}^{N\times N}_{\mathbb{C}}\right)^{|\mathcal{K}|}$ and $c\in\mathbb{R}$ such that

$$\begin{pmatrix}\Psi' & \\ & \eta'\end{pmatrix}=\begin{pmatrix}1 & \\ & P^*\end{pmatrix}\begin{pmatrix}\Psi & \\ & \eta+cI_N\end{pmatrix}\begin{pmatrix}P & \\ & P\end{pmatrix}.$$

Therefore, the equivalence class $[\Psi,\eta]$ is

$$[\Psi,\eta]=\{(\Psi P,P^*(\eta+cI_N)P): P\in\left(\mathcal{O}^{N\times N}_{\mathbb{C}}\right)^{|\mathcal{K}|},\, c\in\mathbb{R}\}.$$

Let $P\in\left(\mathcal{O}^{N\times N}_{\mathbb{C}}\right)^{|\mathcal{K}|}$ and

$$\eta_{\mathrm{k}}=\mathrm{Diag}(\epsilon_{\mathrm{k}1}I_{N_{\mathrm{k}1}},\epsilon_{\mathrm{k}2}I_{N_{\mathrm{k}2}},\ldots,\epsilon_{\mathrm{k}d_{\mathrm{k}}}I_{N_{\mathrm{k}d_{\mathrm{k}}}})_{N\times N},\ \forall \mathrm{k}\in\mathcal{K},$$

then $(\Psi P,\eta)\sim(\Psi,\eta)$ if and only if P_{k} has the same block structure with η_{k} for any $\mathrm{k}\in\mathcal{K}$

$$P_{\mathrm{k}}=\mathrm{Diag}(P_{\mathrm{k}1},P_{\mathrm{k}2},\ldots,P_{\mathrm{k}d_{\mathrm{k}}})_{N\times N},\ P_{\mathrm{k}i}\in\mathcal{O}^{N_{\mathrm{k}i}\times N_{\mathrm{k}i}}_{\mathbb{C}}.$$

If $\eta=(I_N)_{k\in\mathcal{K}}$ is fixed, then $F_{\eta_{\mathrm{k}}}=I_N$ and $(\Psi P,\eta)\sim(\Psi,\eta)$ for any $P\in\left(\mathcal{O}^{N\times N}_{\mathbb{C}}\right)^{|\mathcal{K}|}$, i.e., the energy functional is unitarily invariant. It is nothing but the classical Kohn-Sham DFT model.

Similarly, for the gradient of $\mathcal{F}$, we have the following theorem.

THEOREM 3.2. *Given $c\in\mathbb{R}$, $(\Psi,\eta)\in((H^1_{\#}(\Omega,\mathbb{C}))^N)^{|\mathcal{K}|}\times\left(\mathcal{S}^{N\times N}_{\mathbb{C}}\right)^{|\mathcal{K}|}$, and $P\in\left(\mathcal{O}^{N\times N}_{\mathbb{C}}\right)^{|\mathcal{K}|}$.*

1. *There hold*

$$\begin{aligned}\mathcal{F}_\Psi(\Psi P,P^*(\eta+cI_N)P)&=\mathcal{F}_\Psi(\Psi,\eta)P,\\ \nabla_\Psi\mathcal{F}(\Psi P,P^*(\eta+cI_N)P)&=\nabla_\Psi\mathcal{F}(\Psi,\eta)P,\\ \nabla_\eta\mathcal{F}(\Psi P,P^*(\eta+cI_N)P)&=P^*\nabla_\eta\mathcal{F}(\Psi,\eta)P;\end{aligned}\tag{3.7}$$

2. $\nabla_{\eta_{\mathrm{k}}}\mathcal{F}(\Psi,\eta)$ *is Hermitian matrix for any* $\mathrm{k}\in\mathcal{K}$*;*
3. $\sum_{\mathrm{k}\in\mathcal{K}}\mathrm{tr}\,\nabla_{\eta_{\mathrm{k}}}\mathcal{F}(\Psi,(\eta+cI_N))=0$.

The first property tells us how to apply unitary transformations to Ψ, η and the associated gradients consistently. The third property is the another description of the translation invariance of $\mathcal{F}$ with respect to η and will be used in our convergence analysis.

3.2. Existence of the minimizer. In this subsection, we show the existence of the minimizer of the ensemble Kohn-Sham DFT model. We consider that the sampling of k-points is at Γ point only, for which Ψ, η and other corresponding functions and spaces are of real valued. For the general sampling $\mathcal{K}$, the existence of the minimizer of the ensemble Kohn-Sham DFT model is still open.

Following [6], we assume that $\mathcal{E}_{\rm xc}$ is of the form

$$\mathcal{E}_{\rm xc}(\rho) = \int_\Omega \mathcal{N}(\rho)(r)\mathrm{d}r$$

and

$$\mathcal{N} \in \mathscr{P}(3,(c_1,c_2))\,(c_1 \geq 0) \text{ or } \mathcal{N} \in \mathscr{P}(4/3,(c_1,c_2), \tag{3.8}$$

where

$$\mathscr{P}\left(p,(c_1,c_2)\right) = \{f : \exists a_1, a_2 \in \mathbb{R} \text{ such that } c_1 t^p + a_1 \leq f(t) \leq c_2 t^p + a_2 \quad \forall t \geq 0\}$$

with $c_1 \in \mathbb{R}$ and $p, c_2 \in [0,\infty)$. We assume that there exists a constant $\alpha > 0$ such that for any $\psi \in L^2_{\#}(\Omega)$, the following inequality holds:

$$(\psi, \mathcal{B}\psi) \geq \alpha \|\psi\|^2. \tag{3.9}$$

We also assume that the assumptions A.I-A.IV are satisfied. Let

$$\mathscr{F}_{\rm occ} = \{F = \mathrm{Diag}(f_1, f_2, \ldots, f_N) \in \mathcal{D}^{N\times N} : 2\sum_{i=1}^N f_i = N_e, f_i \in (0,1), i = 1,2,\ldots,N\}.$$

Obviously,

$$\overline{\mathscr{F}}_{\rm occ} = \{F = \mathrm{Diag}(f_1, f_2, \ldots, f_N) \in \mathcal{D}^{N\times N} : 2\sum_{i=1}^N f_i = N_e, f_i \in [0,1], i = 1,2,\ldots,N\}.$$

We first have the following lemma.

LEMMA 3.3. *There holds*

$$\inf_{(\Psi,\eta)\in\mathcal{M}_{\mathcal{B}}^N\times\mathcal{S}^{N\times N}} \mathcal{F}(\Psi,\eta) = \inf_{(\Psi,F)\in\mathcal{M}_{\mathcal{B}}^N\times\mathscr{F}_{\rm occ}} \widetilde{\mathcal{F}}(\Psi,F),$$

where $\widetilde{\mathcal{F}}(\Psi,F) = \widetilde{\mathcal{E}}(\Psi,F) - \sigma\,\mathrm{tr}(S\circ f^{-1})(F)$,

$$\widetilde{\mathcal{E}}(\Psi,F) = \mathrm{tr}\left(\left\langle \Psi^T\left(-\frac{1}{2}\Delta + V_{ext}\right)\Psi\right\rangle F\right) + \mathcal{E}_{\rm HXC}(\tilde{\rho}_{\Psi,F})$$

with $\tilde{\rho}_{\Psi,F} = 2\,\mathrm{tr}((\Psi^T\Psi + \langle\Psi^T M\rangle\mathcal{Q}\langle M^T\Psi\rangle)F)$.

Proof. Let $(\Psi,\eta) \in \mathcal{M}_{\mathcal{B}}^N \times \mathcal{D}^{N\times N}$. We have

$$\mathcal{F}(\Psi,\eta) = \widetilde{\mathcal{F}}(\Psi,F_\eta),$$

which together with (3.5) yields the conclusion. ∎

Let $f(-\infty)=1$, $f(+\infty)=0$ and $S(-\infty)=\lim\limits_{x\to-\infty}S(x)$, $S(+\infty)=\lim\limits_{x\to+\infty}S(x)$, then f and S are continuous on $[-\infty,+\infty]$ and $f([-\infty,\infty])=[0,1]$. Thus $S\circ f^{-1}$ is continuous on $[0,1]$. By Lemma 3.3, instead of $\inf\limits_{(\Psi,F)\in\mathcal{M}_{\mathcal{B}}^N\times\mathcal{S}^{N\times N}}\mathcal{F}(\Psi,\eta)$, we consider the following minimization problem

$$\inf_{(\Psi,F)\in\mathcal{M}_{\mathcal{B}}^N\times\overline{\mathscr{F}}_{\rm occ}}\widetilde{\mathcal{F}}(\Psi,F). \tag{3.10}$$

We shall prove that $\widetilde{\mathcal{F}}$ does indeed have a minimizer on $\mathcal{M}_{\mathcal{B}}^N\times\overline{\mathscr{F}}_{occ}$. Let

$$\widetilde{\widetilde{\mathcal{E}}}(\Psi)=\mathrm{tr}\left(\left\langle\Psi^T\left(-\frac{1}{2}\Delta+V_{\rm ext}\right)\Psi\right\rangle\right)+\frac{1}{2}\int_{\mathbb{R}^3}\frac{\rho_\Psi(r)\rho_\Psi(r')}{|r-r'|}\mathrm{d}r\mathrm{d}r'+\mathcal{E}_{\rm xc}(\rho_\Psi),$$

where $\rho_\Psi=2\,\mathrm{tr}(\Psi^T\Psi+\langle\Psi^TM\rangle\mathcal{Q}\langle M^T\Psi\rangle)$. Then we have

$$\widetilde{\widetilde{\mathcal{E}}}(\Psi F^{1/2})=\widetilde{\mathcal{E}}(\Psi,F),\forall(\Psi,F)\in\mathcal{M}_{\mathcal{B}}^N\times\overline{\mathscr{F}}_{\rm occ}. \tag{3.11}$$

To prove $\widetilde{\mathcal{F}}$ has a minimizer on $\mathcal{M}_{\mathcal{B}}^N\times\overline{\mathscr{F}}_{\rm occ}$, we need the lower semi-continuity of $\widetilde{\widetilde{\mathcal{E}}}$ in the weak topology of $(H^1_{\#}(\Omega))^N$ (See, e.g., [6, 7]).

Proposition 3.4. *Suppose (3.8) holds. If $\Psi^{(n)}$ converges weakly to Ψ in $(H^1_{\#}(\Omega))^N$, then*

$$\widetilde{\widetilde{\mathcal{E}}}(\Psi)\le\varliminf_{n\to\infty}\widetilde{\widetilde{\mathcal{E}}}(\Psi^{(n)}).$$

Using (3.9), Jensen's inequality and the similar arguments in [7], we get that $\widetilde{\mathcal{E}}(\Psi,F)$ is bounded below over $\mathcal{M}_{\mathcal{B}}^N\times\overline{\mathscr{F}}_{\rm occ}$.

Proposition 3.5. *If (3.8) and (3.9) hold, then there exist constants $C>0$ and $b>0$ such that*

$$\widetilde{\mathcal{E}}(\Psi,F)\ge C^{-1}\sum_{i=1}^N\|\Psi F^{1/2}\|^2_{H^1_{\#}}-b\quad\forall(\Psi,F)\in\mathcal{M}_{\mathcal{B}}^N\times\overline{\mathscr{F}}_{\rm occ}.$$

Finally, we obtain the existence of a minimizer for (3.10).

Theorem 3.6. *If (3.8), (3.9) and the assumptions A.I-A.IV hold, then there exits $(\Phi_*,F_*)\in\mathcal{M}_{\mathcal{B}}^N\times\overline{\mathscr{F}}_{\rm occ}$ such that*

$$\widetilde{\mathcal{F}}(\Phi_*,F_*)=\inf_{(\Psi,F)\in\mathcal{M}_{\mathcal{B}}^N\times\overline{\mathscr{F}}_{\rm occ}}\widetilde{\mathcal{F}}(\Psi,F).$$

Proof. Let $\alpha=\inf\limits_{(\Psi,F)\in\mathcal{M}_{\mathcal{B}}^N\times\overline{\mathscr{F}}_{occ}}\widetilde{\mathcal{F}}(\Psi,F)$. It follows from Proposition 3.5 and $S([-\infty,+\infty])$ being bounded that $\alpha>-\infty$. It is clear that $\alpha<\infty$.

Choose $\Psi^{(n)}=(\psi_1^{(n)},\dots,\psi_N^{(n)})\in\mathcal{M}_{\mathcal{B}}^N$ and $F^{(n)}=\mathrm{Diag}(f_1^{(n)},\dots,f_N^{(n)})\in\overline{\mathscr{F}}_{occ}$ such that

$$\lim_{n\to\infty}\widetilde{\mathcal{F}}(\Psi^{(n)},F^{(n)})=\alpha.$$

We then get from Proposition 3.5 that $\Psi^{(n)}(F^{(n)})^{1/2}$ is uniformly bounded in $(H^1_{\#}(\Omega))^N$. We derive from Kakutani's Theorem (see Theorem 4.2 in page 132 of [8]) that there

exists a weakly convergent subsequence of $\Psi^{(n)}(F^{(n)})^{1/2}$ in $(H^1_{\#}(\Omega))^N$. Without loss of generality, let

$$\Psi^{(n)}(F^{(n)})^{1/2} \rightharpoonup \Psi_* = (\psi_{*,1}, \ldots, \psi_{*,N}) \quad \text{in } (H^1_{\#}(\Omega))^N,$$

where $\Psi_* \in (H^1_{\#}(\Omega))^N$. Since $(H^1_{\#}(\Omega))^N$ is compactly embedded into $L^2_{\#}(\Omega)$, we see that $\Psi^{(n)}(F^{(n)})^{1/2} \to \Psi_*$ strongly in $L^2_{\#}(\Omega)$ as $n \to \infty$. Let $F_* = \langle \Psi_*^T \Psi_* \rangle$. We have

$$F^{(n)} = \langle (\Psi^{(n)}(F^{(n)})^{1/2})^T \Psi^{(n)}(F^{(n)})^{1/2} \rangle \to F_*, \tag{3.12}$$

which shows $F_* \in \overline{\mathscr{F}}_{occ}$ and that there exists $\Phi_* \in \mathcal{M}^N_{\mathcal{B}}$ such that $\Phi_* F_*^{1/2} = \Psi_*$. From (3.11), (3.12), and Proposition 3.4, we obtain

$$\begin{aligned}
\widetilde{\mathcal{F}}(\Phi_*, F_*) &= \widetilde{\widetilde{\mathcal{E}}}(\Psi_*(F_*)^{1/2}) - \sigma \operatorname{tr}(S \circ f^{-1})(F_*) \\
&\le \varliminf_{n\to\infty} \widetilde{\widetilde{\mathcal{E}}}(\Psi^{(n)}(F^{(n)})^{1/2}) + \varliminf_{n\to\infty} \left(-\sigma \operatorname{tr}(S \circ f^{-1})(F^{(n)}) \right) \\
&\le \varliminf_{n\to\infty} \left(\widetilde{\widetilde{\mathcal{E}}}(\Psi^{(n)}(F^{(n)})^{1/2}) - \sigma \operatorname{tr}(S \circ f^{-1})(F^{(n)}) \right) \\
&= \varliminf_{n\to\infty} \widetilde{\mathcal{F}}(\Phi^{(n)}, F^{(n)}) \\
&= \alpha.
\end{aligned}$$

This completes the proof. ∎

4. Numerical approximations. We apply the planewave method to discrete (2.6). For any $\mathrm{k} \in \mathcal{K}$, let

$$V_{\mathrm{k},N_G} = \operatorname{span}\left\{ e_{\mathrm{G}} : \mathrm{G} \in \mathcal{R}^*, \frac{1}{2}|\mathrm{k}+\mathrm{G}|^2 \le E_{\mathrm{cut}} \right\},$$

where E_{cut} is a given cutoff energy, N_G is the largest number of planewaves among $\mathrm{k} \in \mathcal{K}$. Consequently, a finite planewave discretization of the ensemble Kohn-Sham DFT minimization problem (2.6) is as follows

$$\inf_{(\Psi,\eta) \in \left(\prod_{\mathrm{k}\in\mathcal{K}} \mathcal{M}^N_{\mathcal{B},\mathbb{C},\mathrm{k},N_G} \right) \times \left(\mathcal{S}^{N\times N}_{\mathbb{C}}\right)^{|\mathcal{K}|}} \mathcal{F}(\Psi, \eta), \tag{4.1}$$

where $\prod$ is the Cartesian product and $\mathcal{M}^N_{\mathcal{B},\mathbb{C},\mathrm{k},N_G}$ is the Stiefel manifold

$$\mathcal{M}^N_{\mathcal{B},\mathbb{C},\mathrm{k},N_G} = \{\Psi \in (V_{\mathrm{k},N_G})^N : \langle \Psi^* \mathcal{B} \Psi \rangle = I_N\}.$$

Since $\prod_{\mathrm{k}\in\mathcal{K}} \mathcal{M}^N_{\mathcal{B},\mathbb{C},\mathrm{k},N_G}$ is compact for any finite sampling, we obtain the existence of a minimizer of the discrete problem (4.1) in the sense of section 3.2. In addition, the invariance of the energy functional and its gradient in section 3.1 also holds since $\prod_{\mathrm{k}\in\mathcal{K}} \mathcal{M}^N_{\mathcal{B},\mathbb{C},\mathrm{k},N_G} \subset ((H^1_{\#}(\Omega,\mathbb{C}))^N)^{|\mathcal{K}|}$.

4.1. Numerical method. We understand that the line search method is widely used to solve a minimization problem, in which there are two main issues: a search direction and a step size. In our minimization problem (4.1), we observe that the iterative behavior for Ψ and η may be different. Hence it is better to apply different step sizes for Ψ and η when we apply the line search method to solve the minimization problem (4.1). Inspired by the adaptive step size strategy proposed in [10], we propose an adaptive double step size strategy for the line search method.

4.1.1. Adaptive double step size strategy. An adaptive step size strategy is concluded as the following four steps [10]:

Initialize $\rightarrow$ **Estimate** $\rightarrow$ **Judge** $\rightarrow$ **Improve**.

We suppose that the initial guess of the step sizes $(t_\Psi^{n,\text{initial}}, t_\eta^{n,\text{initial}})$ at n-th iteration is given. Then we introduce the other three steps of our adaptive double step size strategy one by one.

Let $D_\Psi^{(n)} = (D_{\Psi_{\mathrm{k}}}^{(n)})_{\mathrm{k}\in\mathcal{K}} \in \prod_{\mathrm{k}\in\mathcal{K}} \mathcal{T}_{\Psi_{\mathrm{k}}}\mathcal{M}_{\mathcal{B},\mathbb{C},\mathrm{k},N_G}^N$, $D_\eta^{(n)} = (D_{\eta_{\mathrm{k}}}^{(n)})_{\mathrm{k}\in\mathcal{K}} \in \left(\mathcal{S}_{\mathbb{C}}^{N\times N}\right)^{|\mathcal{K}|}$.

For the sake of convenience, omiting $\Psi^{(n)}, \eta^{(n)}, D_\Psi^{(n)}$ and $D_\eta^{(n)}$, we denote

$$\mathcal{F}((\mathrm{ortho}(\Psi_{\mathrm{k}}^{(n)}, D_{\Psi_{\mathrm{k}}}^{(n)}, t_\Psi))_{\mathrm{k}\in\mathcal{K}}, \eta^{(n)} + t_\eta D_\eta^{(n)})$$

by $\bar{\mathcal{F}}_n(t_\Psi, t_\eta)$, where $\mathrm{ortho}(\Psi_{\mathrm{k}}^{(n)}, D_{\Psi_{\mathrm{k}}}^{(n)}, t_\Psi)$ means one step from $\Psi_{\mathrm{k}}^{(n)} \in \mathcal{M}_{\mathcal{B},\mathbb{C},\mathrm{k},N_G}^N$ with the search direction $D_{\Psi_{\mathrm{k}}}^{(n)}$ and the step size t_Ψ to the next point in $\mathcal{M}_{\mathcal{B},\mathbb{C},\mathrm{k},N_G}^N$. More introduction about $\mathrm{ortho}(\Psi_{\mathrm{k}}^{(n)}, D_{\Psi_{\mathrm{k}}}^{(n)}, t_\Psi)$ will be provided in section 4.1.2. By a simple calculation, we have

$$\frac{\partial\bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0) = 2\,\mathrm{Re}\langle\mathcal{F}_\Psi(\Psi^{(n)},\eta^{(n)}), D_\Psi^{(n)}\rangle,\ \frac{\partial\bar{\mathcal{F}}_n}{\partial t_\eta}(0,0) = \mathrm{Re}\langle\nabla_\eta\mathcal{F}(\Psi^{(n)},\eta^{(n)}), D_\eta^{(n)}\rangle.$$

We assume $\left\langle \left(D_{\Psi_{\mathrm{k}}}^{(n)}\right)^* \mathcal{B}\Psi_{\mathrm{k}}^{(n)} \right\rangle = 0$ for any $\mathrm{k}\in\mathcal{K}$ to ensure

$$\frac{\partial\bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0) = \mathrm{Re}\langle\nabla_\Psi\mathcal{F}(\Psi^{(n)},\eta^{(n)}), D_\Psi^{(n)}\rangle.$$

We always assume that all search directions $D_\Psi^{(n)}$ and $D_\eta^{(n)}$ are descent directions, namely,

$$\frac{\partial\bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0) \le 0,\ \frac{\partial\bar{\mathcal{F}}_n}{\partial t_\eta}(0,0) \le 0, \quad n = 0,1,2,\ldots, \tag{4.2}$$

where $\frac{\partial\bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0) = 0$ if and only if $\nabla_\Psi\mathcal{F}(\Psi^{(n)},\eta^{(n)}) = 0$, and $\frac{\partial\bar{\mathcal{F}}_n}{\partial t_\eta}(0,0) = 0$ if and only if $\nabla_\eta\mathcal{F}(\Psi^{(n)},\eta^{(n)}) = 0$. For simplicity, we always suppose $\|\nabla_\Psi\mathcal{F}(\Psi^{(n)},\eta^{(n)})\| + \|\nabla_\eta\mathcal{F}(\Psi^{(n)},\eta^{(n)})\|_{sF} \neq 0$ in the adaptive double step size strategy, otherwise we have obtained the minimizer of the problem (4.1).

Estimate. The final step sizes are supposed to satisfy the following non-monotone condition:

$$\bar{\mathcal{F}}_n(t_\Psi^{(n)}, t_\eta^{(n)}) - \mathcal{C}_n \le \nu\left(t_\Psi^{(n)}\frac{\partial\bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0) + t_\eta^{(n)}\frac{\partial\bar{\mathcal{F}}_n}{\partial t_\eta}(0,0)\right),\ n = 0,1,2,\ldots, \tag{4.3}$$

where $\nu\in(0,1)$ is a given parameter. Here $\mathcal{C}_n$ can be $\mathcal{F}(\Psi^{(n)},\eta^{(n)})$ or that introduced in [39] as follows

$$\begin{cases} \mathcal{C}_0 = \mathcal{F}(\Psi^{(0)},\eta^{(0)}),\ Q_0 = 1, \\ Q_n = \alpha Q_{n-1} + 1, \\ \mathcal{C}_n = (\alpha Q_{n-1}\mathcal{C}_{n-1} + \mathcal{F}(\Psi^{(n)},\eta^{(n)}))/Q_n, \end{cases} \tag{4.4}$$

where $\alpha\in[0,1)$ is a given parameter. We consider the approximation of the energy functional $\mathcal{F}$ around $(\Psi^{(n)},\eta^{(n)})$ as follows:

$$\bar{\mathcal{F}}_n(t_\Psi,t_\eta) \approx \bar{\mathcal{F}}_n(0,0) + t_\Psi\frac{\partial\bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0) + t_\eta\frac{\partial\bar{\mathcal{F}}_n}{\partial t_\eta}(0,0) + \frac{1}{2}c_{n,1}t_\Psi^2 + \frac{1}{2}c_{n,2}t_\eta^2, \tag{4.5}$$

where $c_{n,1}, c_{n,2} \geq 0$ are approximations of the second derivatives, $c_{n,1} = 0$ if and only if $\nabla_\Psi \mathcal{F}(\Psi^{(n)}, \eta^{(n)}) = 0$, and $c_{n,2} = 0$ if and only if $\nabla_\eta \mathcal{F}(\Psi^{(n)}, \eta^{(n)}) = 0$. Replacing $\bar{\mathcal{F}}_n(t_\Psi^{(n)}, t_\eta^{(n)})$ in (4.3) by the right hand term of (4.5), we obtain

$$\begin{aligned}
&\bar{\mathcal{F}}_n(0,0) + t_\Psi \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0) + t_\eta \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\eta}(0,0) + \frac{1}{2} c_{n,1} t_\Psi^2 + \frac{1}{2} c_{n,2} t_\eta^2 - \mathcal{C}_n \\
&\leq \nu \left(t_\Psi \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0) + t_\eta \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\eta}(0,0) \right),
\end{aligned}$$

or equivalently,

$$\frac{\bar{\mathcal{F}}_n(0,0) + t_\Psi \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0) + t_\eta \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\eta}(0,0) + \frac{1}{2} c_{n,1} t_\Psi^2 + \frac{1}{2} c_{n,2} t_\eta^2 - \mathcal{C}_n}{t_\Psi \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0) + t_\eta \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\eta}(0,0)} \geq \nu.$$

Hence, we propose the following estimator

$$\zeta_n(t_\Psi, t_\eta) = \frac{\bar{\mathcal{F}}_n(0,0) + t_\Psi \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0) + t_\eta \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\eta}(0,0) + \frac{1}{2} c_{n,1} t_\Psi^2 + \frac{1}{2} c_{n,2} t_\eta^2 - \mathcal{C}_n}{t_\Psi \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0) + t_\eta \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\eta}(0,0)} \tag{4.6}$$

to guide us whether to accept the step sizes or not at the n-th iteration. Since the estimator (4.6) remains reliable only in a neighborhood of $(\Psi^{(n)}, \eta^{(n)})$, it is reasonable to restrict $t_\Psi^{(n)} \|D_\Psi^{(n)}\|_\infty \leq \theta_\Psi^{(n)}$ and $t_\eta^{(n)} \|D_\eta^{(n)}\|_{sF,\infty} \leq \theta_\eta^{(n)}$ for some given small $\theta_\Psi^{(n)}$, $\theta_\eta^{(n)} \in (0,1)$. Thus, we first set

$$t_\Psi^{(n)} = \min\left(t_\Psi^{n,\text{initial}}, \frac{\theta_\Psi^{(n)}}{\|D_\Psi^{(n)}\|_\infty} \right), \; t_\eta^{(n)} = \min\left(t_\eta^{n,\text{initial}}, \frac{\theta_\eta^{(n)}}{\|D_\eta^{(n)}\|_{sF,\infty}} \right),$$

and then calculate the estimator $\zeta_n(t_\Psi^{(n)}, t_\eta^{(n)})$.

Judge. The estimator $\zeta_n(t_\Psi^{(n)}, t_\eta^{(n)})$ is used to determine whether to accept the step sizes $(t_\Psi^{(n)}, t_\eta^{(n)})$ or not. If $(t_\Psi^{(n)}, t_\eta^{(n)})$ satisfies

$$\zeta_n(t_\Psi^{(n)}, t_\eta^{(n)}) \geq \nu, \tag{4.7}$$

then we accept this step sizes. Otherwise, $(t_\Psi^{(n)}, t_\eta^{(n)})$ is to be improved.

Improve. If $(t_\Psi^{(n)}, t_\eta^{(n)})$ is not accepted , then we solve the minimizer of the approximation (4.5) of $\bar{\mathcal{F}}_n$ and set it to be the step size. Combining the restriction of approximation in the neighborhood of $(\Psi^{(n)}, \eta^{(n)})$, we take

$$\begin{aligned}
t_\Psi^{(n)} &= \min\left(-\frac{1}{c_{n,1}} \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0), \frac{\theta_\Psi^{(n)}}{\|D_\Psi^{(n)}\|_\infty} \right), \\
t_\eta^{(n)} &= \min\left(-\frac{1}{c_{n,2}} \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\eta}(0,0), \frac{\theta_\eta^{(n)}}{\|D_\eta^{(n)}\|_{sF,\infty}} \right).
\end{aligned} \tag{4.8}$$

Here and hereafter, $-\frac{1}{c_{n,1}} \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0)$ is replaced by $-\frac{1}{c_{n,2}} \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\eta}(0,0)$ if $\nabla_\Psi \mathcal{F}(\Psi^{(n)}, \eta^{(n)}) = 0$, and $-\frac{1}{c_{n,2}} \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\eta}(0,0)$ is replaced by $-\frac{1}{c_{n,1}} \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0)$ if $\nabla_\eta \mathcal{F}(\Psi^{(n)}, \eta^{(n)}) = 0$. Note that we choose $\nu \in (0, 1/2]$ to ensure that step sizes (4.8) satisfy (4.7). To ensure the

convergence of the iterations, we may do some adjustments on the above step sizes. More precisely, if

$$\underline{c} \le \frac{t_\eta^{(n)}}{t_\Psi^{(n)}} \le \bar{c} \tag{4.9}$$

does not hold, we then reduce one of two step sizes to make them satisfy the above inequalities. Here $\bar{c} > 1 > \underline{c} > 0$ are given constants.

REMARK 4.1. *We can always choose $c_{n,1}, c_{n,2}$ such that the minimizer of (4.5) satisfies $t_\Psi = t_\eta$, i.e.,*

$$-\frac{1}{c_{n,1}}\frac{\partial \bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0) = -\frac{1}{c_{n,2}}\frac{\partial \bar{\mathcal{F}}_n}{\partial t_\eta}(0,0).$$

In this case, the minimizer of (4.5) is also the minimizer of the following function

$$\bar{\mathcal{F}}_n(0,0) + \left(\frac{\partial \bar{\mathcal{F}}_n}{\partial t}(0,0) + \frac{\partial \bar{\mathcal{F}}_n}{\partial t}(0,0)\right) t + \frac{1}{2}(c_{n,1} + c_{n,2})t^2.$$

Hence, the approximation of $\bar{\mathcal{F}}_n$ with the same step size $t_\Psi = t_\eta$ is a special case of the above discussion.

We summarize the above process as Algorithm 4.1.

Note that it is very difficult to calculate the second derivatives of $\bar{\mathcal{F}}_n(t_\Psi, t_\eta)$. Thus we design some strategies to get good approximations $c_{n,1}$ and $c_{n,2}$. We provide three strategies to get $c_{n,1}$ and $c_{n,2}$ by one trial step with step sizes $(t_\Psi^{\text{trial}}, t_\eta^{\text{trial}})$. For convenience, we use the short notation

$$\tilde{\mathcal{F}}_n(t_\Psi, t_\eta) = \bar{\mathcal{F}}_n(0,0) + t_\Psi \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0) + t_\eta \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\eta}(0,0) + \frac{1}{2}c_{n,1}t_\Psi^2 + \frac{1}{2}c_{n,2}t_\eta^2.$$

We shall also simply denote $\bar{\mathcal{F}}_n(t,t)$ and $\tilde{\mathcal{F}}_n(t,t)$ by $\bar{\mathcal{F}}_n(t)$ and $\tilde{\mathcal{F}}_n(t)$, respectively. In this case, $c_{n,1} + c_{n,2}$ are denoted by c_n, trial step sizes t_Ψ^{trial} and t_η^{trial} are denoted by t^{trial}.

(S1) Applying the same step size $t_\Psi = t_\eta$ for Ψ and η, we use the energy at t^{trial} to get the approximation $\tilde{\mathcal{F}}_n$, namely, $\tilde{\mathcal{F}}_n$ satisfies

$$\tilde{\mathcal{F}}_n(t^{\text{trial}}) = \bar{\mathcal{F}}_n(t^{\text{trial}}),$$

where

$$t^{\text{trial}} = \min\left(\max\left(t^{\min}, t^{(n-1)}\right), \frac{\theta^{(n)}}{\sqrt{\|D_\Psi^{(n)}\|_\infty^2 + \|D_\eta^{(n)}\|_{sF,\infty}^2}}\right),$$

$t^{\min}$ and $\theta^{(n)} \in (0,1)$ are given parameters. Then we have

$$c_n = \frac{2(\bar{\mathcal{F}}_n(t^{\text{trial}}) - \bar{\mathcal{F}}_n(0) - t^{\text{trial}}\bar{\mathcal{F}}_n'(0))}{(t^{\text{trial}})^2}.$$

We choose

$$t_\Psi^{(n)} = t_\eta^{(n)} = \begin{cases} \min\left(t_m^{(n)}, \frac{\theta^{(n)}}{\sqrt{\|D_\Psi^{(n)}\|_\infty^2 + \|D_\eta^{(n)}\|_{sF,\infty}^2}}\right), & t_m^{(n)} > 0, \\ t^{\text{trial}}, & \text{otherwise}, \end{cases}$$

Algorithm 4.1 Adaptive double step size strategy

Input: Ψ, η, D_Ψ, D_η, t_Ψ^{initial}, t_η^{initial}, $t_\Psi^{\min}$, $t_\eta^{\min}$, ν, c_1, c_2, θ_Ψ, θ_η, $\mathcal{C}$

1: Set

$$t_\Psi = \min\left(\max(t_\Psi^{\text{initial}}, t_\Psi^{\min}), \frac{\theta_\Psi}{\|D_\Psi\|_\infty}\right),$$

$$t_\eta = \min\left(\max(t_\eta^{\text{initial}}, t_\eta^{\min}), \frac{\theta_\eta}{\|D_\eta\|_{sF,\infty}}\right);$$

2: Calculate the estimator

$$\zeta(t_\Psi, t_\eta) = \frac{\bar{\mathcal{F}}(0,0) + t_\Psi \frac{\partial \bar{\mathcal{F}}}{\partial t_\Psi}(0,0) + t_\eta \frac{\partial \bar{\mathcal{F}}}{\partial t_\eta}(0,0) + \frac{1}{2}c_1 t_\Psi^2 + \frac{1}{2}c_2 t_\eta^2 - \mathcal{C}}{t_\Psi \frac{\partial \bar{\mathcal{F}}}{\partial t_\Psi}(0,0) + t_\eta \frac{\partial \bar{\mathcal{F}}}{\partial t_\eta}(0,0)},$$

where $\bar{\mathcal{F}}(t_\Psi, t_\eta) = \mathcal{F}((\text{ortho}(\Psi_{\text{k}}, D_{\Psi_{\text{k}}}, t_\Psi))_{\text{k}\in\mathcal{K}}, \eta + t_\eta D_\eta)$;

3: **if** $\zeta(t_\Psi, t_\eta) < \nu$ **then**

4: set

$$t_\Psi = \min\left(-\frac{1}{c_1}\frac{\partial \bar{\mathcal{F}}}{\partial t_\Psi}(0,0), \frac{\theta_\Psi}{\|D_\Psi\|_\infty}\right),$$

$$t_\eta = \min\left(-\frac{1}{c_2}\frac{\partial \bar{\mathcal{F}}}{\partial t_\eta}(0,0), \frac{\theta_\eta}{\|D_\eta\|_{sF,\infty}}\right);$$

5: **end if**

6: **if** $\frac{t_\eta}{t_\Psi} < \underline{c}$ **then**

7: $t_\Psi = \frac{1}{\underline{c}} t_\eta,\ t_\eta = t_\eta$;

8: **else if** $\frac{t_\eta}{t_\Psi} > \bar{c}$ **then**

9: $t_\Psi = t_\Psi,\ t_\eta = \bar{c} t_\Psi$;

10: **end if**

11: Return (t_Ψ, t_η).

where

$$t_m^{(n)} = -\frac{\bar{\mathcal{F}}_n'(0)}{c_n} = -\frac{\bar{\mathcal{F}}_n'(0)(t^{\text{trial}})^2}{2(\bar{\mathcal{F}}_n(t^{\text{trial}}) - \bar{\mathcal{F}}_n(0) - \bar{\mathcal{F}}_n'(0)t^{\text{trial}})}.$$

(S2) Applying the same step size $t_\Psi = t_\eta$ for Ψ and η, we use the derivative of $\bar{\mathcal{F}}_n(t)$ at t^{trial} to get the approximation $\tilde{\mathcal{F}}_n$, namely, $\tilde{\mathcal{F}}_n$ satisfies

$$\tilde{\mathcal{F}}_n'(t^{\text{trial}}) = \bar{\mathcal{F}}_n'(t^{\text{trial}}),$$

where

$$t^{\text{trial}} = \min\left(\max(t^{\min}, t^{(n-1)}), \frac{\theta^{(n)}}{\sqrt{\|D_\Psi^{(n)}\|_\infty^2 + \|D_\eta^{(n)}\|_{sF,\infty}^2}}\right),$$

$t^{\min}$ and $\theta^{(n)} \in (0,1)$ are given parameters. Then we have

$$c_n = \frac{\bar{\mathcal{F}}_n'(t^{\text{trial}}) - \bar{\mathcal{F}}_n'(0)}{t^{\text{trial}}}.$$

We choose

$$t_\Psi^{(n)} = t_\eta^{(n)} = \begin{cases} \min\left(t_m^{(n)}, \frac{\theta^{(n)}}{\sqrt{\|D_\Psi^{(n)}\|_\infty^2 + \|D_\eta^{(n)}\|_{sF,\infty}^2}}\right), & t_m^{(n)} > 0, \\ t^{\text{trial}}, & \text{otherwise}, \end{cases}$$

where

$$t_m^{(n)} = -\frac{\bar{\mathcal{F}}_n'(0)}{c_n} = -\frac{\bar{\mathcal{F}}_n'(0) t^{\text{trial}}}{\bar{\mathcal{F}}_n'(t^{\text{trial}}) - \mathcal{F}_n'(0)}.$$

(S3) Applying different step sizes $t_\Psi \neq t_\eta$ for Ψ and η, we use partial derivatives of $\bar{\mathcal{F}}_n(t_\Psi, t_\eta)$ at $(t_\Psi^{\text{trial}}, t_\eta^{\text{trial}})$ to get the approximation $\tilde{\mathcal{F}}_n$, namely, $\tilde{\mathcal{F}}_n$ satisfies

$$\frac{\partial \tilde{\mathcal{F}}_n}{\partial t_\Psi}(t_\Psi^{\text{trial}}, t_\eta^{\text{trial}}) = \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\Psi}(t_\Psi^{\text{trial}}, t_\eta^{\text{trial}}), \quad \frac{\partial \tilde{\mathcal{F}}_n}{\partial t_\eta}(t_\Psi^{\text{trial}}, t_\eta^{\text{trial}}) = \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\eta}(t_\Psi^{\text{trial}}, t_\eta^{\text{trial}}),$$

where

$$t_\Psi^{\text{trial}} = \min\left(\max(t^{\min,\Psi}, t_\Psi^{(n-1)}), \frac{\theta_\Psi^{(n)}}{\|D_\Psi^{(n)}\|_\infty}\right),$$
$$t_\eta^{\text{trial}} = \min\left(\max(t^{\min,\eta}, t_\eta^{(n-1)}), \frac{\theta_\eta^{(n)}}{\|D_\eta^{(n)}\|_{sF,\infty}}\right),$$

$(t_\Psi^{\min}, t_\eta^{\min})$ and $\theta_\Psi^{(n)}, \theta_\eta^{(n)} \in (0,1)$ are given parameters. Then we have

$$c_{n,1} = \frac{\frac{\partial \bar{\mathcal{F}}_n}{\partial t_\Psi}(t_\Psi^{\text{trial}}, t_\eta^{\text{trial}}) - \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0)}{t_\Psi^{\text{trial}}}, \quad c_{n,2} = \frac{\frac{\partial \bar{\mathcal{F}}_n}{\partial t_\eta}(t_\Psi^{\text{trial}}, t_\eta^{\text{trial}}) - \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\eta}(0,0)}{t_\eta^{\text{trial}}}.$$

We choose

$$\begin{cases} t_\Psi^{(n)} = \min\left(t_{m,\Psi}^{(n)}, \frac{\theta_\Psi^{(n)}}{\|D_\Psi^{(n)}\|_\infty}\right),\ t_\eta^{(n)} = \min\left(t_{m,\eta}^{(n)}, \frac{\theta_\eta^{(n)}}{\|D_\eta^{(n)}\|_{sF,\infty}}\right), & t_{m,\Psi}^{(n)} > 0 \text{ and } t_{m,\eta}^{(n)} > 0, \\ t_\Psi^{(n)} = t_\Psi^{\text{trial}},\ t_\eta^{(n)} = t_\eta^{\text{trial}}, & \text{otherwise}, \end{cases}$$

where

$$t_{m,\Psi}^{(n)} = -\frac{\frac{\partial \bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0)}{c_{n,1}} = -\frac{\frac{\partial \bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0) t_\Psi^{\text{trial}}}{\frac{\partial \bar{\mathcal{F}}_n}{\partial t_\Psi}(t_\Psi^{\text{trial}}, t_\eta^{\text{trial}}) - \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0)},$$
$$t_{m,\eta}^{(n)} = -\frac{\frac{\partial \bar{\mathcal{F}}_n}{\partial t_\eta}(0,0)}{c_{n,2}} = -\frac{\frac{\partial \bar{\mathcal{F}}_n}{\partial t_\eta}(0,0) t_\eta^{\text{trial}}}{\frac{\partial \bar{\mathcal{F}}_n}{\partial t_\eta}(t_\Psi^{\text{trial}}, t_\eta^{\text{trial}}) - \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\eta}(0,0)}.$$

For strategies (S2) and (S3), we need to calculate the following two partial derivatives

$$\frac{\partial \bar{\mathcal{F}}_n}{\partial t_\Psi}(t_\Psi^{\text{trial}}, t_\eta^{\text{trial}}), \ \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\eta}(t_\Psi^{\text{trial}}, t_\eta^{\text{trial}}).$$

A direct calculation shows

$$\frac{\partial \bar{\mathcal{F}}_n}{\partial t_\Psi}(t_\Psi^{\text{trial}}, t_\eta^{\text{trial}})$$
$$= \left\langle \mathcal{F}_\Psi((\text{ortho}(\Psi_{\text{k}}^{(n)}, D_{\Psi_{\text{k}}}^{(n)}, t_\Psi^{\text{trial}}))_{\text{k}\in\mathcal{K}}, \eta^{(n)} + t_\eta^{\text{trial}} D_\eta^{(n)}), \left(\frac{\partial\, \text{ortho}(\Psi_{\text{k}}^{(n)}, D_{\Psi_{\text{k}}}^{(n)}, t_\Psi^{\text{trial}})}{\partial t}\right)_{\text{k}\in\mathcal{K}} \right\rangle.$$

and

$$\frac{\partial \bar{\mathcal{F}}_n}{\partial t_\eta}(t_\Psi^{\text{trial}}, t_\eta^{\text{trial}}) = \left\langle \nabla_\eta \mathcal{F}((\text{ortho}(\Psi_{\text{k}}^{(n)}, D_{\Psi_{\text{k}}}^{(n)}, t_\Psi^{\text{trial}}))_{\text{k}\in\mathcal{K}}, \eta^{(n)} + t_\eta^{\text{trial}} D_\eta^{(n)}), D_\eta^{(n)} \right\rangle.$$

We see that $\dfrac{\partial \,\text{ortho}(\Psi_{\text{k}}^{(n)}, D_{\Psi_{\text{k}}}^{(n)}, t_\Psi^{\text{trial}})}{\partial t}$ is very difficult to calculate. Instead, we apply the third order approximation

$$\begin{aligned}\frac{\partial \,\text{ortho}(\Psi_{\text{k}}^{(n)}, D_{\Psi_{\text{k}}}^{(n)}, t_\Psi^{\text{trial}})}{\partial t} \approx{}& \frac{\partial \,\text{ortho}(\Psi_{\text{k}}^{(n)}, D_{\Psi_{\text{k}}}^{(n)}, 0)}{\partial t} + \frac{\partial^2 \,\text{ortho}(\Psi_{\text{k}}^{(n)}, D_{\Psi_{\text{k}}}^{(n)}, 0)}{\partial t^2} t_\Psi^{\text{trial}} \\ &+ \frac{1}{2}\frac{\partial^3 \,\text{ortho}(\Psi_{\text{k}}^{(n)}, D_{\Psi_{\text{k}}}^{(n)}, 0)}{\partial t^3}(t_\Psi^{\text{trial}})^2\end{aligned}$$

in practice.

4.1.2. The preconditioned conjugate gradient method. Now we introduce the preconditioned conjugate gradient method for solving the minimization problem (4.1). The preconditioned conjugate gradient (PCG) method is a typical line search based optimization method. For the constrained optimization problem (4.1), we usually need to keep each iteration point on the constrained manifold. Thus some unitarity preserving strategies are required. We then introduce the preconditioner, the conjugate gradient parameter and the unitarity preserving strategies one by one.

We first introduce the preconditioner applied to $\nabla_\Psi \mathcal{F}$ and $\nabla_\eta \mathcal{F}$. Let $(\Psi, \eta) \in \left(\prod_{\text{k}\in\mathcal{K}} \mathcal{M}^N_{\mathcal{B},\mathbb{C},\text{k},N_G}\right) \times \left(\mathcal{S}_{\mathbb{C}}^{N\times N}\right)^{|\mathcal{K}|}$, where all η_{k} are diagonal matrices. We consider a preconditioner in the form of $M_\Psi^\eta(\Phi) = (M_{\Psi_{\text{k}}}^{\eta_{\text{k}}}(\Phi_{\text{k}}))_{\text{k}\in\mathcal{K}}$ for $\nabla_\Psi \mathcal{F}$, where

$$M_{\Psi_{\text{k}}}^{\eta_{\text{k}}}(\Phi_{\text{k}}) = M_{\Psi_{\text{k}}}\left(\frac{1}{2w_{\text{k}}}\Phi_{\text{k}} F_{\eta_{\text{k}}}^{-1}\right)$$

and $M_{\Psi_{\text{k}}} : V_{\text{k},N_G} \to V_{\text{k},N_G}$ is a linear operator. In our numerical experiments, we apply the following preconditioner $M_{\Psi_{\text{k}}}$ used in Quantum ESPRESSO [33]

$$[M_{\Psi_{\text{k}}}]_{G,G'} = \delta_{G,G'} \frac{1}{1 + \frac{1}{2}|\text{k}+G|^2 + \sqrt{1 + \left(\frac{1}{2}|\text{k}+G|^2 - 1\right)^2}},$$

which is independent of wavefunctions. We consider a preconditioner in the form of $M_\eta(A) = (M_{\eta_{\text{k}}}(A_{\text{k}}))_{\text{k}\in\mathcal{K}}$ for $\nabla_\eta \mathcal{F}$, where $M_{\eta_{\text{k}}} : \mathcal{S}_{\mathbb{C}}^{N\times N} \to \mathcal{S}_{\mathbb{C}}^{N\times N}$ is a linear operator defined by

$$(M_{\eta_{\text{k}}}(A_{\text{k}}))_{ij} = -A_{\text{k}ij}\frac{1}{w_{\text{k}}}\frac{\eta_{\text{k}ii} - \eta_{\text{k}jj}}{f_{\text{k}j} - f_{\text{k}i}},\ \forall i,j = 1,2,\ldots,N,\ \forall \text{k} \in \mathcal{K}. \tag{4.10}$$

Here $f_{\text{k}i} = f((\eta_{\text{k}ii} - \mu)/\sigma)$ and

$$\frac{f_{\text{k}j} - f_{\text{k}i}}{\eta_{\text{k}ii} - \eta_{\text{k}jj}} = \frac{1}{\sigma} f'\left(\frac{\eta_{\text{k}ii} - \mu}{\sigma}\right)$$

when $\eta_{\text{k}ii} = \eta_{\text{k}jj}$.

Applying $M_{\Psi_{\text{k}}}^{\eta_{\text{k}}}$ to

$$\nabla_{\Psi_{\text{k}}} \mathcal{F}(\Psi, \eta) = 2w_{\text{k}}(H_{\text{k}}(\rho_{\Psi,\eta})\Psi - \mathcal{B}\Psi_{\text{k}}\Sigma_{\text{k}})F_{\eta_{\text{k}}},$$

we obtain

$$M_{\Psi_{\rm k}}^{\eta_{\rm k}}(\nabla_{\Psi_{\rm k}}\mathcal{F}(\Psi,\eta)) = M_{\Psi_{\rm k}}(H_{\rm k}(\rho_{\Psi,\eta})\Psi - \mathcal{B}\Psi_{\rm k}\Sigma_{\rm k}).$$

Here $\Sigma_{\rm k} = \langle \Psi_{\rm k}^* H_{\rm k}(\rho_{\Psi,\eta})\Psi_{\rm k}\rangle$. Compared to $\nabla_{\Psi_{\rm k}}\mathcal{F}(\Psi,\eta)$, $M_{\Psi_{\rm k}}^{\eta_{\rm k}}(\nabla_{\Psi_{\rm k}}\mathcal{F}(\Psi,\eta))$ eliminates the occupation number $F_{\eta_{\rm k}}$ and $2w_{\rm k}$. We see that

$$(\nabla_{\Psi_{\rm k}}\mathcal{F}(\Psi,\eta))_i = 2w_{\rm k}(H_{\rm k}(\rho_{\Psi,\eta})\psi_{{\rm k}i} - (\mathcal{B}\Psi_{\rm k}\Sigma_{\rm k})_i)(F_{\eta_{\rm k}})_{ii}$$

is almost 0 when the occupation number $(F_{\eta_{\rm k}})_{ii}$ is close to 0. Consequently, the preconditioner $M_{\Psi_{\rm k}}^{\eta_{\rm k}}$ removes $F_{\eta_{\rm k}}$ in $\nabla_{\Psi_{\rm k}}\mathcal{F}(\Psi,\eta)$ to eliminate the impact of small occupation numbers on the convergence rate, which has been mentioned in [21, 28].

Applying $M_{\eta_{\rm k}}$ to $\nabla_{\eta_{\rm k}}\mathcal{F}(\Psi,\eta)$, we have

$$M_{\eta_{\rm k}}(\nabla_{\eta_{\rm k}}\mathcal{F}(\Psi,\eta)) = cI + \eta_{\rm k} - \Sigma_{\rm k},$$

where c is defined by (B.3). We note that $\kappa(\eta_{\rm k} - \Sigma_{\rm k})$ is the preconditioned gradient mentioned in [14], where κ is some positive constant.

We then introduce the conjugate gradient parameters. The typical choices of the conjugate gradient parameters include the Hestenes-Stiefel (HS) formula [20], the Polak-Ribiére-Polyak (PRP) formula [31, 32], the Fletcher-Reeves (FR) formula [13] and the Dai-Yuan (DY) formula [11]. In our numerical experiments, we choose the DY formula, which is expressed as

$$\beta^{(n)} = \frac{{\rm Re}\left(\left\langle M_{\Psi^{(n)}}^{\eta^{(n)}}(G_\Psi^{(n)}), G_\Psi^{(n)}\right\rangle + \left\langle M_{\eta^{(n)}}(G_\eta^{(n)}), G_\eta^{(n)}\right\rangle\right)}{{\rm Re}\left(\left\langle D_\Psi^{(n-1)}, G_\Psi^{(n)} - G_\Psi^{(n-1)}\right\rangle + \left\langle D_\eta^{(n-1)}, G_\eta^{(n)} - G_\eta^{(n-1)}\right\rangle\right)}$$

for the PCG algorithm, where Re gives the real part, $G_\Psi^{(n)} = \nabla_\Psi\mathcal{F}(\Psi^{(n)},\eta^{(n)})$, $G_\eta^{(n)} = \nabla_\eta\mathcal{F}(\Psi^{(n)},\eta^{(n)})$. Hereafter, we shall sometimes use the notations $G_\Psi^{(n)}$ and $G_\eta^{(n)}$ to simplify some formulas.

Now we turn to introduce the unitarity preserving strategy we use. Let $D_{\Psi_{\rm k}} \in \mathcal{T}_{\Psi_{\rm k}}\mathcal{M}_{\mathcal{B},\mathbb{C},{\rm k},N_G}^N$. We denote by

$$\mathrm{ortho}(\Psi_{\rm k}, D_{\Psi_{\rm k}}, t_\Psi)$$

one step from $\Psi_{\rm k} \in \mathcal{M}_{\mathcal{B},\mathbb{C},{\rm k},N_G}^N$ with the search direction $D_{\Psi_{\rm k}}$ and the step size t_Ψ to the next point in $\mathcal{M}_{\mathcal{B},\mathbb{C},{\rm k},N_G}^N$. In our numerical experiments, we apply the QR strategy, which is defined by

$$\mathrm{ortho}_{\rm QR}(\Psi_{\rm k}, D_{\Psi_{\rm k}}, t_\Psi) = (\Psi_{\rm k} + t_\Psi D_{\Psi_{\rm k}})L^{-*}, \tag{4.11}$$

where L is the lower triangular matrix such that

$$LL^* = I_N + t_\Psi^2\langle D_{\Psi_{\rm k}}^*\mathcal{B}D_{\Psi_{\rm k}}\rangle.$$

We refer [9] for some other unitarity preserving strategies such as the PD strategy.

We assume $\mathrm{ortho}(\Psi_{\rm k}, D_{\Psi_{\rm k}}, t_\Psi)$ satisfies the following assumption, which is needed in our analysis and valid for both QR and PD strategy (see, e.g., [9]).

Assumption 4.2. *There exist constants $C_1, C_2 > 0$ such that*

$$\|\,\mathrm{ortho}(\Phi, D_\Phi, t) - \Phi\| \le C_1 t\|D_\Phi\|, \quad \forall t \ge 0,$$
$$\left\|\frac{\partial}{\partial t}\,\mathrm{ortho}(\Phi, D_\Phi, t) - D_\Phi\right\| \le C_2 t\|D_\Phi\|^2, \quad \forall t \ge 0$$

for any $\Phi \in \mathcal{M}_{\mathcal{B}}^N$ and $D_\Phi \in \mathcal{T}_\Phi\mathcal{M}_{\mathcal{B}}^N$.

We now propose our preconditioned conjugate gradient method as Algorithm 4.2.

Algorithm 4.2 PCG method

1: Given $\alpha \in [0,1)$, $\nu \in (0,1/2]$, $t_\Psi^{\min}, t_\eta^{\min}, E_{\text{cut}} > 0$, and choose the initial data $\Psi_{\mathrm{k}}^{(0)} \in \mathcal{M}_{\mathcal{B},\mathbb{C},\mathrm{k},N_G}^N$ and $\eta_{\mathrm{k}}^{(0)} = \mathrm{Diag}(\epsilon_{\mathrm{k}1}^{(0)}, \ldots, \epsilon_{\mathrm{k}N}^{(0)})$ for any $\mathrm{k} \in \mathcal{K}$. Let $D_\Psi^{(-1)} = (D_{\Psi_{\mathrm{k}}}^{(-1)})_{\mathrm{k}\in\mathcal{K}} = 0$, $D_\eta^{(-1)} = (D_{\eta_{\mathrm{k}}}^{(-1)})_{\mathrm{k}\in\mathcal{K}} = 0$, $n = 0$;

2: Calculate the gradient $G_\Psi^{(n)} = (G_{\Psi_{\mathrm{k}}}^{(n)})_{\mathrm{k}\in\mathcal{K}}$, $G_\eta^{(n)} = (G_{\eta_{\mathrm{k}}}^{(n)})_{\mathrm{k}\in\mathcal{K}}$ and the preconditioned gradient $\widetilde{G}_\Psi^{(n)} = M_{\Psi^{(n)}}^{\eta^{(n)}}(G_\Psi^{(n)})$, $\widetilde{G}_\eta^{(n)} = M_{\eta^{(n)}}(G_\eta^{(n)})$, where $G_{\Psi_{\mathrm{k}}}^{(n)} = \nabla_{\Psi_{\mathrm{k}}}\mathcal{F}(\Psi^{(n)}, \eta^{(n)})$, $G_{\eta_{\mathrm{k}}}^{(n)} = \nabla_{\eta_{\mathrm{k}}}\mathcal{F}(\Psi^{(n)}, \eta^{(n)})$

3: Calculate the conjugate gradient parameter $\beta^{(n)}$;

4: Calculate the search direction

$$D_\Psi^{(n)} = -\widetilde{G}_\Psi^{(n)} + \beta^{(n)} D_\Psi^{(n-1)}, \ D_\eta^{(n)} = -\widetilde{G}_\eta^{(n)} + \beta^{(n)} D_\eta^{(n-1)};$$

5: Project the search direction $D_{\Psi_{\mathrm{k}}}^{(n)}$ to the tangent space $\mathcal{T}_{\Psi_{\mathrm{k}}}\mathcal{M}_{\mathcal{B},\mathbb{C},\mathrm{k},N_{\mathrm{G}}}^N$

$$D_{\Psi_{\mathrm{k}}}^{(n)} = P_{0,\Psi_{\mathrm{k}}^{(n)}}^*(D_{\Psi_{\mathrm{k}}}^{(n)}), \ \forall \mathrm{k} \in \mathcal{K};$$

6: Set $D_{\Psi_{\mathrm{k}}}^{(n)} = -D_{\Psi_{\mathrm{k}}}^{(n)} \operatorname{sign} \mathrm{Re}\left\langle G_\Psi^{(n)}, D_\Psi^{(n)} \right\rangle$, $D_{\eta_{\mathrm{k}}}^{(n)} = -D_{\eta_{\mathrm{k}}}^{(n)} \operatorname{sign} \mathrm{Re}\left\langle G_\eta^{(n)}, D_\eta^{(n)} \right\rangle$ for any $\mathrm{k} \in \mathcal{K}$;

7: Choose the appropriate parameters $(\theta_\Psi^{(n)}, \theta_\eta^{(n)})$;

8: Calculate $\mathcal{C}_n$ by (4.4);

9: Given the initial guess of the step sizes $(t_\Psi^{n,\text{initial}}, t_\eta^{n,\text{initial}})$;

10: Give $c_{n,1}$ and $c_{n,2}$ and calculate $t_\Psi^{(n)}$ and $t_\eta^{(n)}$ by

$$\begin{aligned}&(t_\Psi^{(n)}, t_\eta^{(n)})\\ &= \text{Adaptve double step size strategy}(\Psi^{(n)}, \eta^{(n)}, D_\Psi^n, D_\eta^{(n)}, t_\Psi^{n,\text{initial}}, t_\eta^{n,\text{initial}},\\ &\qquad\qquad t_\Psi^{\min}, t_\eta^{\min}, \nu, c_{n,1}, c_{n,2}, \theta_\Psi^{(n)}, \theta_\eta^{(n)}, \mathcal{C}_n);\end{aligned}$$

11: Set $\Psi_{\mathrm{k}}^{(n+1)} = \mathrm{ortho}(\Psi_{\mathrm{k}}^{(n)}, D_{\Psi_{\mathrm{k}}}^{(n)}, t_\Psi^{(n)})$, $\eta_{\mathrm{k}}^{(n+1)} = \eta_{\mathrm{k}}^{(n)} + t_\eta^{(n)} D_{\eta_{\mathrm{k}}}^{(n)}$ for any $\mathrm{k} \in \mathcal{K}$;

12: Pick up $P^{(n+1)} = (P_{\mathrm{k}}^{(n+1)})_{\mathrm{k}\in\mathcal{K}} \in (\mathcal{O}_{\mathbb{C}}^{N\times N})^{|\mathcal{K}|}$ such that $(P_{\mathrm{k}}^{(n+1)})^* \eta_{\mathrm{k}}^{(n+1)} P_{\mathrm{k}}^{(n+1)}$ is diagonal for any $\mathrm{k} \in \mathcal{K}$ and then update

$$\begin{aligned}\Psi^{(n+1)} &= \Psi^{(n+1)} P^{(n+1)}, \quad \eta^{(n+1)} = (P^{(n+1)})^* \eta^{(n+1)} P^{(n+1)},\\ D_\Psi^{(n)} &= D_\Psi^{(n)} P^{(n+1)}, \quad D_\eta^{(n)} = (P^{(n+1)})^* D_\eta^{(n)} P^{(n+1)};\end{aligned}$$

13: Let $n = n+1$. Convergence check: if not converged, go to step 2; else, stop.

We see that $D_\Psi^{(n)}$ in the 4-th step of Algorithm 4.2 is not in the tangent space $\prod_{\mathrm{k}\in\mathcal{K}} \mathcal{T}_{\Psi_{\mathrm{k}}^{(n)}}\mathcal{M}_{\mathcal{B},\mathbb{C},\mathrm{k},N_G}^N$. Thus we project $D_{\Psi_{\mathrm{k}}}^{(n)}$ to $\mathcal{T}_{\Psi_{\mathrm{k}}^{(n)}}\mathcal{M}_{\mathcal{B},\mathbb{C},\mathrm{k},N_G}^N$ in the 5-th step. In

order to ensure

$$\frac{\partial \bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0) = \mathrm{Re}\langle \nabla_\Psi \mathcal{F}(\Psi^{(n)}, \eta^{(n)}), D_\Psi^{(n)} \rangle,$$

we apply the projection $P^*_{0,\Psi_{\mathrm{k}}^{(n)}}$ for each $\mathrm{k} \in \mathcal{K}$.

4.1.3. The restarted preconditioned conjugate gradient method. To get better approximations, we turn to consider the restarted preconditioned conjugate gradient method.

In practice, we expect that there exists a positive constant a such that

$$\overline{\lim_{n\to\infty}} \frac{-\mathrm{Re}\left(\left\langle G_\Psi^{(n)}, D_\Psi^{(n)} \right\rangle + \left\langle G_\eta^{(n)}, D_\eta^{(n)} \right\rangle\right)}{\left|\left\langle G_\Psi^{(n)}, M_{\Psi^{(n)}}^{\eta^{(n)}}(G_\Psi^{(n)}) \right\rangle\right|^a + \left|\left\langle G_\eta^{(n)}, M_{\eta^{(n)}}(G_\eta^{(n)}) \right\rangle\right|^a} > 0. \tag{4.12}$$

Here $G_\Psi^{(n)} = \nabla_\Psi \mathcal{F}(\Psi^{(n)}, \eta^{(n)})$ and $G_\eta^{(n)} = \nabla_\eta \mathcal{F}(\Psi^{(n)}, \eta^{(n)})$. Thus we restart the PCG method when

$$\frac{-\mathrm{Re}\left(\left\langle G_\Psi^{(n)}, D_\Psi^{(n)} \right\rangle + \left\langle G_\eta^{(n)}, D_\eta^{(n)} \right\rangle\right)}{\left|\left\langle G_\Psi^{(n)}, M_{\Psi^{(n)}}^{\eta^{(n)}}(G_\Psi^{(n)}) \right\rangle\right|^a + \left|\left\langle G_\eta^{(n)}, M_{\eta^{(n)}}(G_\eta^{(n)}) \right\rangle\right|^a} < \gamma, \tag{4.13}$$

for some given parameter $\gamma \in (0,1)$. Applying this strategy, we propose a restarted preconditioned conjugate gradient method shown as Algorithm 4.3.

In the numerical experiments, we observe that retarting directly is sometimes better than changing the sign of the search direction when the preconditioned conjugate gradient direction is not a descent direction. Thus we propose a new restarted preconditioned conjugate gradient method shown as Algorithm 4.4.

4.2. Convergence analysis. In this subsection, we analyze the convergence of the restarted PCG methods (Algorithms 4.3 and 4.4). For convenience, we show the detailed proofs for the case that the sampling of k-points is at Γ point only. For the general sampling $\mathcal{K}$, the convergence of the restarted PCG method can be obtained by the similar arguments. We shall sometimes use the notations $G_\Psi^{(n)} = \nabla_\Psi \mathcal{F}(\Psi^{(n)}, \eta^{(n)})$ and $G_\eta^{(n)} = \nabla_\eta \mathcal{F}(\Psi^{(n)}, \eta^{(n)})$ to simplify some formulas.

We first give some assumptions which is needed in our analysis.

ASSUMPTION 4.3. *There exist $\alpha_\Psi, \alpha_\eta > 0$ such that*

$$\begin{aligned} \langle \nabla_\Psi \mathcal{F}(\Psi,\eta), M_\Psi^\eta(\nabla_\Psi \mathcal{F}(\Psi,\eta)) \rangle &\geq \alpha_\Psi \|\nabla_\Psi \mathcal{F}(\Psi,\eta)\|^2, \\ \langle \nabla_\eta \mathcal{F}(\Psi,\eta), M_\eta(\nabla_\eta \mathcal{F}(\Psi,\eta)) \rangle &\geq \alpha_\eta \|\nabla_\eta \mathcal{F}(\Psi,\eta)\|_{sF}^2 \end{aligned} \tag{4.14}$$

for $(\Psi,\eta) \in \mathcal{M}_{\mathcal{B},N_G}^N \times \mathcal{S}^{N\times N}$.

We obtain from the assumption above that the preconditioner is bounded from below uniformly. We see that M_Ψ^η we applied always satisfies (4.14) and M_η we applied satisfies (4.14) when f is strictly monotonically decreasing.

ASSUMPTION 4.4. *The gradient of $\mathcal{F}$ is Lipschitz continuous. That is, there exists $L_0 > 0$ such that*

$$\begin{aligned} &\|\mathcal{F}_\Psi(\Psi_1,\eta_1) - \mathcal{F}_\Psi(\Psi_2,\eta_2)\| + \|\mathcal{F}_\eta(\Psi_1,\eta_1) - \mathcal{F}_\eta(\Psi_2,\eta_2)\|_{sF} \\ &\leq L_0(\|\Psi_1 - \Psi_2\| + \|\eta_1 - \eta_2\|_{sF}) \end{aligned}$$

for any $(\Psi_1,\eta_1), (\Psi_2,\eta_2) \in \mathcal{M}_{\mathcal{B},N_G}^N \times \mathcal{S}^{N\times N}$.

Algorithm 4.3 Restarted PCG method I

1: Given $\alpha \in [0,1)$, $\nu \in (0,1/2]$, $a, t_\Psi^{\min}, t_\eta^{\min}, E_{\rm cut} > 0$, and choose the initial data $\Psi_{\rm k}^{(0)} \in \mathcal{M}^N_{\mathcal{B},\mathbb{C},{\rm k},N_G}$ and $\eta_{\rm k}^{(0)} = {\rm Diag}(\epsilon_{{\rm k}1}^{(0)}, \ldots, \epsilon_{{\rm k}N}^{(0)})$ for any ${\rm k} \in \mathcal{K}$. Let $D_\Psi^{(-1)} = (D_{\Psi_{\rm k}}^{(-1)})_{{\rm k}\in\mathcal{K}} = 0$, $D_\eta^{(-1)} = (D_{\eta_{\rm k}}^{(-1)})_{{\rm k}\in\mathcal{K}} = 0$, $n = 0$;

2: Calculate the gradient $G_\Psi^{(n)} = (G_{\Psi_{\rm k}}^{(n)})_{{\rm k}\in\mathcal{K}}$, $G_\eta^{(n)} = (G_{\eta_{\rm k}}^{(n)})_{{\rm k}\in\mathcal{K}}$ and the preconditioned gradient $\widetilde{G}_\Psi^{(n)} = M_{\Psi^{(n)}}^{\eta^{(n)}}(G_\Psi^{(n)})$, $\widetilde{G}_\eta^{(n)} = M_{\eta^{(n)}}(G_\eta^{(n)})$, where $G_{\Psi_{\rm k}}^{(n)} = \nabla_{\Psi_{\rm k}}\mathcal{F}(\Psi^{(n)}, \eta^{(n)})$, $G_{\eta_{\rm k}}^{(n)} = \nabla_{\eta_{\rm k}}\mathcal{F}(\Psi^{(n)}, \eta^{(n)})$

3: Calculate the conjugate gradient parameter $\beta^{(n)}$;

4: Calculate the search direction

$$D_\Psi^{(n)} = -\widetilde{G}_\Psi^{(n)} + \beta^{(n)} D_\Psi^{(n-1)}, \ D_\eta^{(n)} = -\widetilde{G}_\eta^{(n)} + \beta^{(n)} D_\eta^{(n-1)};$$

5: Project the search direction $D_{\Psi_{\rm k}}^{(n)}$ to the tangent space $\mathcal{T}_{\Psi_{\rm k}}\mathcal{M}^N_{\mathcal{B},\mathbb{C},{\rm k},N_{\rm G}}$

$$D_{\Psi_{\rm k}}^{(n)} = P^*_{0,\Psi_{\rm k}^{(n)}}(D_{\Psi_{\rm k}}^{(n)}), \ \forall {\rm k} \in \mathcal{K};$$

6: Set $D_{\Psi_{\rm k}}^{(n)} = -D_{\Psi_{\rm k}}^{(n)} \, {\rm sign}\, {\rm Re}\left\langle G_\Psi^{(n)}, D_\Psi^{(n)} \right\rangle$, $D_{\eta_{\rm k}}^{(n)} = -D_{\eta_{\rm k}}^{(n)} \, {\rm sign}\, {\rm Re}\left\langle G_\eta^{(n)}, D_\eta^{(n)} \right\rangle$ for any ${\rm k} \in \mathcal{K}$;

7: **if** (4.13) holds **then**

8: $\quad D_{\Psi_{\rm k}}^{(n)} = -P^*_{0,\Psi_{\rm k}^{(n)}}(\widetilde{G}_{\Psi_{\rm k}}^{(n)}), \ D_{\eta_{\rm k}}^{(n)} = -\widetilde{G}_{\eta_{\rm k}}^{(n)}, \ \forall {\rm k} \in \mathcal{K}$;

9: **end if**

10: Choose the appropriate parameters $(\theta_\Psi^{(n)}, \theta_\eta^{(n)})$;

11: Calculate $\mathcal{C}_n$ by (4.4);

12: Given the initial guess of the step sizes $(t_\Psi^{n,{\rm initial}}, t_\eta^{n,{\rm initial}})$;

13: Give $c_{n,1}$ and $c_{n,2}$ and calculate $t_\Psi^{(n)}$ and $t_\eta^{(n)}$ by

$$\begin{aligned}&(t_\Psi^{(n)}, t_\eta^{(n)})\\ &= \text{Adaptve double step size strategy}(\Psi^{(n)}, \eta^{(n)}, D_\Psi^n, D_\eta^{(n)}, t_\Psi^{n,{\rm initial}}, t_\eta^{n,{\rm initial}},\\ &\qquad\qquad t_\Psi^{\min}, t_\eta^{\min}, \nu, c_{n,1}, c_{n,2}, \theta_\Psi^{(n)}, \theta_\eta^{(n)}, \mathcal{C}_n);\end{aligned}$$

14: Set $\Psi_{\rm k}^{(n+1)} = {\rm ortho}(\Psi_{\rm k}^{(n)}, D_{\Psi_{\rm k}}^{(n)}, t_\Psi^{(n)})$, $\eta_{\rm k}^{(n+1)} = \eta_{\rm k}^{(n)} + t_\eta^{(n)} D_{\eta_{\rm k}}^{(n)}$ for any ${\rm k} \in \mathcal{K}$;

15: Pick up $P^{(n+1)} = (P_{\rm k}^{(n+1)})_{{\rm k}\in\mathcal{K}} \in (\mathcal{O}_{\mathbb{C}}^{N\times N})^{|\mathcal{K}|}$ such that $(P_{\rm k}^{(n+1)})^* \eta_{\rm k}^{(n+1)} P_{\rm k}^{(n+1)}$ is diagonal for any ${\rm k} \in \mathcal{K}$ and then update

$$\begin{aligned}\Psi^{(n+1)} &= \Psi^{(n+1)} P^{(n+1)}, \quad \eta^{(n+1)} = (P^{(n+1)})^* \eta^{(n+1)} P^{(n+1)},\\ D_\Psi^{(n)} &= D_\Psi^{(n)} P^{(n+1)}, \quad D_\eta^{(n)} = (P^{(n+1)})^* D_\eta^{(n)} P^{(n+1)};\end{aligned}$$

16: Let $n = n+1$. Convergence check: if not converged, go to step 2; else, stop.

Algorithm 4.4 Restarted PCG method II

1: Given $\alpha \in [0,1)$, $\nu \in (0,1/2]$, $a, t_\Psi^{\min}, t_\eta^{\min}, E_{\rm cut} > 0$, and choose the initial data $\Psi_{\rm k}^{(0)} \in \mathcal{M}_{\mathcal{B},\mathbb{C},{\rm k},N_G}^N$ and $\eta_{\rm k}^{(0)} = {\rm Diag}(\epsilon_{{\rm k}1}^{(0)}, \ldots, \epsilon_{{\rm k}N}^{(0)})$ for any ${\rm k} \in \mathcal{K}$. Let $D_\Psi^{(-1)} = (D_{\Psi_{\rm k}}^{(-1)})_{{\rm k}\in\mathcal{K}} = 0$, $D_\eta^{(-1)} = (D_{\eta_{\rm k}}^{(-1)})_{{\rm k}\in\mathcal{K}} = 0$, $n = 0$;

2: Calculate the gradient $G_\Psi^{(n)} = (G_{\Psi_{\rm k}}^{(n)})_{{\rm k}\in\mathcal{K}}$, $G_\eta^{(n)} = (G_{\eta_{\rm k}}^{(n)})_{{\rm k}\in\mathcal{K}}$ and the preconditioned gradient $\widetilde{G}_\Psi^{(n)} = M_{\Psi^{(n)}}^{\eta^{(n)}}(G_\Psi^{(n)})$, $\widetilde{G}_\eta^{(n)} = M_{\eta^{(n)}}(G_\eta^{(n)})$, where $G_{\Psi_{\rm k}}^{(n)} = \nabla_{\Psi_{\rm k}}\mathcal{F}(\Psi^{(n)}, \eta^{(n)})$, $G_{\eta_{\rm k}}^{(n)} = \nabla_{\eta_{\rm k}}\mathcal{F}(\Psi^{(n)}, \eta^{(n)})$

3: Calculate the conjugate gradient parameter $\beta^{(n)}$;

4: Calculate the search direction

$$D_\Psi^{(n)} = -\widetilde{G}_\Psi^{(n)} + \beta^{(n)} D_\Psi^{(n-1)}, \ D_\eta^{(n)} = -\widetilde{G}_\eta^{(n)} + \beta^{(n)} D_\eta^{(n-1)};$$

5: Project the search direction $D_{\Psi_{\rm k}}^{(n)}$ to the tangent space $\mathcal{T}_{\Psi_{\rm k}}\mathcal{M}_{\mathcal{B},\mathbb{C},{\rm k},N_{\rm G}}^N$

$$D_{\Psi_{\rm k}}^{(n)} = P_{0,\Psi_{\rm k}^{(n)}}^*(D_{\Psi_{\rm k}}^{(n)}), \ \forall {\rm k} \in \mathcal{K};$$

6: **if** $\operatorname{sign}\operatorname{Re}\left\langle G_\Psi^{(n)}, D_\Psi^{(n)} \right\rangle \geq 0$ or $\operatorname{sign}\operatorname{Re}\left\langle G_\eta^{(n)}, D_\eta^{(n)} \right\rangle \geq 0$ or (4.13) holds **then**

7: $\quad D_{\Psi_{\rm k}}^{(n)} = -P_{0,\Psi_{\rm k}^{(n)}}^*(\widetilde{G}_{\Psi_{\rm k}}^{(n)}), \ D_{\eta_{\rm k}}^{(n)} = -\widetilde{G}_{\eta_{\rm k}}^{(n)}, \ \forall {\rm k} \in \mathcal{K};$

8: **end if**

9: Choose the appropriate parameters $(\theta_\Psi^{(n)}, \theta_\eta^{(n)})$;

10: Calculate $\mathcal{C}_n$ by (4.4);

11: Given the initial guess of the step sizes $(t_\Psi^{n,\rm initial}, t_\eta^{n,\rm initial})$;

12: Give $c_{n,1}$ and $c_{n,2}$ and calculate $t_\Psi^{(n)}$ and $t_\eta^{(n)}$ by

$$\begin{aligned}&(t_\Psi^{(n)}, t_\eta^{(n)})\\ &= \text{Adaptve double step size strategy}(\Psi^{(n)}, \eta^{(n)}, D_\Psi^n, D_\eta^{(n)}, t_\Psi^{n,\rm initial}, t_\eta^{n,\rm initial},\\ &\qquad t_\Psi^{\min}, t_\eta^{\min}, \nu, c_{n,1}, c_{n,2}, \theta_\Psi^{(n)}, \theta_\eta^{(n)}, \mathcal{C}_n);\end{aligned}$$

13: Set $\Psi_{\rm k}^{(n+1)} = {\rm ortho}(\Psi_{\rm k}^{(n)}, D_{\Psi_{\rm k}}^{(n)}, t_\Psi^{(n)})$, $\eta_{\rm k}^{(n+1)} = \eta_{\rm k}^{(n)} + t_\eta^{(n)} D_{\eta_{\rm k}}^{(n)}$ for any ${\rm k} \in \mathcal{K}$;

14: Pick up $P^{(n+1)} = (P_{\rm k}^{(n+1)})_{{\rm k}\in\mathcal{K}} \in (\mathcal{O}_{\mathbb{C}}^{N\times N})^{|\mathcal{K}|}$ such that $(P_{\rm k}^{(n+1)})^* \eta_{\rm k}^{(n+1)} P_{\rm k}^{(n+1)}$ is diagonal for any ${\rm k} \in \mathcal{K}$ and then update

$$\begin{aligned}\Psi^{(n+1)} &= \Psi^{(n+1)} P^{(n+1)}, \quad \eta^{(n+1)} = (P^{(n+1)})^* \eta^{(n+1)} P^{(n+1)},\\ D_\Psi^{(n)} &= D_\Psi^{(n)} P^{(n+1)}, \quad D_\eta^{(n)} = (P^{(n+1)})^* D_\eta^{(n)} P^{(n+1)};\end{aligned}$$

15: Let $n = n+1$. Convergence check: if not converged, go to step 2; else, stop.

ASSUMPTION 4.5. *There exists a constant $\bar{C} > 0$ such that*

$$c_{n,1} + c_{n,2} \leq \bar{C}(\|D_\Psi^{(n)}\|^2 + \|D_\eta^{(n)}\|_{sF}^2), \quad n = 0, 1, 2, \ldots \tag{4.15}$$

ASSUMPTION 4.6. *There holds*

$$\underline{c} \leq \frac{-\frac{1}{c_{n,2}}\frac{\partial \bar{\mathcal{F}}_n}{\partial t_\eta}(0,0)}{-\frac{1}{c_{n,1}}\frac{\partial \bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0)} \leq \bar{c}, \quad n = 0, 1, 2, \ldots \tag{4.16}$$

We observe that the assumption 4.5 is similar to that the Hessian of $\mathcal{F}$ is bounded. If the same step sizes for Ψ and η are applied, then we see from Remark 4.1 that Assumption 4.6 is satisfied. And we can always choose some $c_{n,1}$ and $c_{n,2}$ such that Assumptions 4.5 and 4.6 hold.

ASSUMPTION 4.7. *For the subsequence $\{n_j\}_{j\in\mathbb{N}}$ satisfying*

$$\lim_{j\to\infty} \frac{-\left(\left\langle G_\Psi^{(n_j)}, D_\Psi^{(n_j)}\right\rangle + \left\langle G_\eta^{(n_j)}, D_\eta^{(n_j)}\right\rangle\right)}{\left|\left\langle G_\Psi^{(n_j)}, M_{\Psi^{(n_j)}}^{\eta^{(n_j)}}(G_\Psi^{(n_j)})\right\rangle\right|^a + \left|\left\langle G_\eta^{(n_j)}, M_{\eta^{(n_j)}}(G_\eta^{(n_j)})\right\rangle\right|^a} \neq 0,$$

there exists a constant $C > 0$ such that

$$\|D_\Psi^{(n_j)}\| + \|D_\eta^{(n_j)}\| \leq C, \ \forall j \in \mathbb{N}. \tag{4.17}$$

We see that the above assumption can be satisfied by many strategies in practice. For example, if the preconditioned gradients in the iterations are bounded uniformly, we can restart the algorithm when the conjugate gradient parameter is very large. Then we obtain uniformly bounded search directions.

In the following lemma, we need the following assumption for the step sizes.

$$\varliminf_{n\to\infty} t_\Psi^{(n)} > 0, \quad \varliminf_{n\to\infty} t_\eta^{(n)} > 0. \tag{4.18}$$

LEMMA 4.8. *Suppose Assumption 4.3 holds and the sequence $\{(\Psi^{(n)}, \eta^{(n)})\}_{n\in\mathbb{N}}$ is generated by Algorithm 4.2. If $D_\Psi^{(n)}$ and $D_\eta^{(n)}$ satisfy (4.2) and (4.12), $t_\Psi^{(n)}$ and $t_\eta^{(n)}$ satisfy (4.3) and (4.18), then either*

$$\|\nabla_\Psi \mathcal{F}(\Psi^n, \eta^{(n)})\| = 0, \ \|\nabla_\eta \mathcal{F}(\Psi^n, \eta^{(n)})\|_{sF} = 0$$

for some positive n or

$$\varliminf_{n\to\infty} (\|\nabla_\Psi \mathcal{F}(\Psi^n, \eta^{(n)})\| + \|\nabla_\eta \mathcal{F}(\Psi^n, \eta^{(n)})\|_{sF}) = 0.$$

Proof. Suppose

$$\|\nabla_\Psi \mathcal{F}(\Psi^n, \eta^{(n)})\| + \|\nabla_\eta \mathcal{F}(\Psi^n, \eta^{(n)})\|_{sF} \neq 0, \forall n \in \mathbb{N},$$

otherwise the conclusion is true. It follows from the definition of $\mathcal{C}_n$ that for any $n \geq 1$, there holds

$$\mathcal{F}(\Psi^{(n+1)}, \eta^{(n+1)}) - \mathcal{F}(\Psi^{(n)}, \eta^{(n)}) = \mathcal{F}(\Psi^{(n+1)}, \eta^{(n+1)}) - \mathcal{C}_n - \frac{\alpha Q_{n-1}}{Q_n}(\mathcal{F}(\Psi^{(n)}, \eta^{(n)}) - \mathcal{C}_{n-1}).$$

Since

$$\mathcal{F}(\Psi^{(1)}, \eta^{(1)}) - \mathcal{F}(\Psi^{(0)}, \eta^{(0)}) = \mathcal{F}(\Psi^{(1)}, \eta^{(1)}) - \mathcal{C}_0,$$

summing up all $n \in \mathbb{N}$ gives that

$$\begin{aligned}
&\sum_{n=0}^{\infty}(\mathcal{F}(\Psi^n,\eta^{(n)})-\mathcal{F}(\Psi^{(n+1)},\eta^{(n+1)}))\\
=&-\sum_{n=0}^{\infty}(\mathcal{F}(\Psi^{n+1},\eta^{(n+1)})-\mathcal{C}_n)+\sum_{n=0}^{\infty}\frac{\alpha Q_n}{Q_{n+1}}(\mathcal{F}(\Psi^{(n+1)},\eta^{(n+1)})-\mathcal{C}_n)\\
=&-\sum_{n=0}^{\infty}\frac{1}{Q_{n+1}}(\mathcal{F}(\Psi^{n+1},\eta^{(n+1)})-\mathcal{C}_n)\\
\geq&-\nu\sum_{n=0}^{\infty}\frac{1}{Q_{n+1}}\left(t_{\Psi}^{(n)}\frac{\partial\bar{\mathcal{F}}_n}{\partial t_{\Psi}}(0,0)+t_{\eta}^{(n)}\frac{\partial\bar{\mathcal{F}}_n}{\partial t_{\eta}}(0,0)\right).
\end{aligned}$$

Note that $Q_n = 1+\sum_{i=1}^{n}\alpha^i \in [1,\frac{1}{1-\alpha}]$, which together with (4.2) leads to

$$-\sum_{n=0}^{\infty}t_{\Psi}^{(n)}\frac{\partial\bar{\mathcal{F}}_n}{\partial t_{\Psi}}(0,0)<+\infty,\quad -\sum_{n=0}^{\infty}t_{\eta}^{(n)}\frac{\partial\bar{\mathcal{F}}_n}{\partial t_{\eta}}(0,0)<+\infty.$$

Hence

$$\lim_{n\to\infty}t_{\Psi}^{(n)}\frac{\partial\bar{\mathcal{F}}_n}{\partial t_{\Psi}}(0,0)=0,\quad \lim_{n\to\infty}t_{\eta}^{(n)}\frac{\partial\bar{\mathcal{F}}_n}{\partial t_{\eta}}(0,0)=0.$$

Then by (4.18), we have

$$\lim_{n\to\infty}\frac{\partial\bar{\mathcal{F}}_n}{\partial t_{\Psi}}(0,0)=0,\quad \lim_{n\to\infty}\frac{\partial\bar{\mathcal{F}}_n}{\partial t_{\eta}}(0,0)=0,$$

which arrive at

$$\lim_{n\to\infty}\left(-\frac{\partial\bar{\mathcal{F}}_n}{\partial t_{\Psi}}(0,0)-\frac{\partial\bar{\mathcal{F}}_n}{\partial t_{\eta}}(0,0)\right)=0.$$

Since $-\frac{\partial\bar{\mathcal{F}}_n}{\partial t_{\Psi}}(0,0)-\frac{\partial\bar{\mathcal{F}}_n}{\partial t_{\eta}}(0,0)$ is a product of

$$\left|\left\langle G_{\Psi}^{(n)},M_{\Psi^{(n)}}^{\eta^{(n)}}(G_{\Psi}^{(n)})\right\rangle\right|^a+\left|\left\langle G_{\eta}^{(n)},M_{\eta^{(n)}}(G_{\eta}^{(n)})\right\rangle\right|^a$$

and

$$\frac{-\left(\left\langle G_{\Psi}^{(n)},D_{\Psi}^{(n)}\right\rangle+\left\langle G_{\Psi}^{(n)},D_{\eta}^{(n)}\right\rangle\right)}{\left|\left\langle G_{\Psi}^{(n)},M_{\Psi^{(n)}}^{\eta^{(n)}}(G_{\Psi}^{(n)})\right\rangle\right|^a+\left|\left\langle G_{\eta}^{(n)},M_{\eta^{(n)}}(G_{\eta}^{(n)})\right\rangle\right|^a},$$

we obtain from (4.12) that

$$\varliminf_{n\to\infty}\left|\left\langle G_{\Psi}^{(n)},M_{\Psi^{(n)}}^{\eta^{(n)}}(G_{\Psi}^{(n)})\right\rangle\right|^a+\left|\left\langle G_{\eta}^{(n)},M_{\eta^{(n)}}(G_{\eta}^{(n)})\right\rangle\right|^a=0.$$

Consequently, we get from (4.14) that

$$\varliminf_{n\to\infty}(\|\nabla_{\Psi}\mathcal{F}(\Psi^n,\eta^{(n)})\|+\|\nabla_{\eta}\mathcal{F}(\Psi^n,\eta^{(n)})\|_{sF})=0,$$

which completes the proof. ∎

REMARK 4.9. *We see from the above proof that we may only need to consider the subsequence* $\{n_j\}_{j\in\mathbb{N}}$ *satisfying*

$$\lim_{j\to\infty}\frac{-\left(\left\langle G_\Psi^{(n_j)}, D_\Psi^{(n_j)}\right\rangle + \left\langle G_\eta^{(n_j)}, D_\eta^{(n_j)}\right\rangle\right)}{\left|\left\langle G_\Psi^{(n_j)}, M_{\Psi^{(n_j)}}^{\eta^{(n_j)}}(G_\Psi^{(n_j)})\right\rangle\right|^a + \left|\left\langle G_\eta^{(n_j)}, M_{\eta^{(n_j)}}(G_\eta^{(n_j)})\right\rangle\right|^a} \neq 0.$$

In addition, (4.18) *can be replaced by that* (4.9) *holds for the above* $\{n_j\}_{j\in\mathbb{N}}$ *and*

$$\sum_{j=0}^{\infty} t_\Psi^{(n_j)} = +\infty. \tag{4.19}$$

We mention that (4.19) *is weaker than* (4.18) *under the premise of* (4.9).

THEOREM 4.10. *Suppose* $\mathcal{F}$ *is continuously differentiable in* $H^1_{\#}(\Omega)\times\mathcal{S}^{N\times N}$ *and* $\mathcal{F}_\Psi$ *is bounded, i.e., there exists* $C_0>0$ *such that*

$$\|\mathcal{F}_\Psi(\Psi,\eta)\|\le C_0,\ \forall(\Psi,\eta)\in\mathcal{M}^N_{\mathcal{B},N_G}\times\mathcal{S}^{N\times N}, \tag{4.20}$$

and Assumptions 4.2 - 4.6 hold true. Let $\{(D_\Psi^{(n)}, D_\eta^{(n)})\}_{n\in\mathbb{N}}$ *and* $\{(\Psi^{(n)},\eta^{(n)})\}_{n\in\mathbb{N}}$ *are generated by Algorithm 4.3 or Algorithm 4.4. If* $\{(D_\Psi^{(n)}, D_\eta^{(n)})\}_{n\in\mathbb{N}}$ *satisfies Assumption 4.7, then there exists a positive sequence* $\{(\theta_\Psi^{(n)},\theta_\eta^{(n)})\}_{n\in\mathbb{N}}$ *such that either*

$$\|\nabla_\Psi\mathcal{F}(\Psi^n,\eta^{(n)})\|=0,\ \|\nabla_\eta\mathcal{F}(\Psi^n,\eta^{(n)})\|_{sF}=0$$

for some $n>0$ *or*

$$\varliminf_{n\to\infty}\left(\|\nabla_\Psi\mathcal{F}(\Psi^n,\eta^{(n)})\|+\|\nabla_\eta\mathcal{F}(\Psi^n,\eta^{(n)})\|_{sF}\right)=0.$$

Proof. Let

$$\begin{aligned}(\theta_\Psi^{(n)},\theta_\eta^{(n)}) = \sup\Big\{&(\tilde\theta_\Psi^{(n)},\tilde\theta_\eta^{(n)}) : \bar{\mathcal{F}}_n(t_\Psi,t_\eta)-\bar{\mathcal{F}}_n(0,0)-t_\Psi\frac{\partial\bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0)-t_\eta\frac{\partial\bar{\mathcal{F}}_n}{\partial t_\eta}(0,0)\\
&-\frac{1}{2}c_{n,1}t_\Psi^2-\frac{1}{2}c_{n,2}t_\eta^2\le-\frac{\nu}{2}\left(t_\Psi^{(n)}\frac{\partial\bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0)+t_\eta^{(n)}\frac{\partial\bar{\mathcal{F}}_n}{\partial t_\eta}(0,0)\right)\\
&\text{for any } (t_\Psi,t_\eta)\in T_{\tilde\theta_\Psi^{(n)},\tilde\theta_\eta^{(n)}},\ \underline{c}\le\frac{\tilde\theta_\eta^{(n)}}{\|D_\eta^{(n)}\|_{sF}}\Big/\frac{\tilde\theta_\Psi^{(n)}}{\|D_\Psi^{(n)}\|}\le\bar c\\
&\text{when } \|D_\Psi^{(n)}\|\neq0 \text{ and } \|D_\eta^{(n)}\|_{sF}\neq0,\\
&\tilde\theta_\Psi^{(n)}=1 \text{ when } \|D_\Psi^{(n)}\|\neq0, \text{ and } \tilde\theta_\eta^{(n)}=1 \text{ when } \|D_\eta^{(n)}\|_{sF}=0\Big\},\end{aligned}$$

where sup is in the sense of lexicographical order and

$$T_{\tilde\theta_\Psi^{(n)},\tilde\theta_\eta^{(n)}}=\left\{(t_\Psi,t_\eta):0\le t_\Psi\le\frac{\tilde\theta_\Psi^{(n)}}{\|D_\Psi^{(n)}\|},\ 0\le t_\eta\le\frac{\tilde\theta_\eta^{(n)}}{\|D_\eta^{(n)}\|_{sF}},\ \text{and } \underline{c}\le\frac{t_\eta}{t_\Psi}\le\bar c\right\}.$$

Then we prove that the conclusion is valid when above $(\theta_\Psi^{(n)},\theta_\eta^{(n)})$ are taken.

Suppose

$$\|\nabla_\Psi\mathcal{F}(\Psi^n,\eta^{(n)})\|+\|\nabla_\eta\mathcal{F}(\Psi^n,\eta^{(n)})\|_{sF}\neq0,\ \forall n\in\mathbb{N},$$

otherwise the conclusion is true. In Algorithm 4.3 or Algorithm 4.4, it follows from Assumption 4.6 that every $t_\Psi^{(n)}$ and $t_\eta^{(n)}$ satisfies

$$\zeta_n(t_\Psi^{(n)}, t_\eta^{(n)}) \geq \nu,$$
$$t_\Psi^{(n)}\|D_\Psi^{(n)}\| \leq \theta_\Psi^{(n)},\ t_\eta^{(n)}\|D_\eta^{(n)}\|_{sF} \leq \theta_\eta^{(n)},$$

which implies

$$\begin{aligned}&\bar{\mathcal{F}}_n(0,0) + t_\Psi^{(n)}\frac{\partial\bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0) + t_\eta^{(n)}\frac{\partial\bar{\mathcal{F}}_n}{\partial t_\eta}(0,0) + \frac{1}{2}c_{n,1}(t_\Psi^{(n)})^2 + \frac{1}{2}c_{n,2}(t_\eta^{(n)})^2 - \mathcal{C}_n\\ &\leq \nu\left(t_\Psi^{(n)}\frac{\partial\bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0) + t_\eta^{(n)}\frac{\partial\bar{\mathcal{F}}_n}{\partial t_\eta}(0,0)\right).\end{aligned}$$

Then we obtain from the definition of $(\theta_\Psi^{(n)}, \theta_\eta^{(n)})$ that

$$\bar{\mathcal{F}}_n(t_\Psi^{(n)}, t_\eta^{(n)}) - \mathcal{C}_n \leq \frac{\nu}{2}\left(t_\Psi^{(n)}\frac{\partial\bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0) + t_\eta^{(n)}\frac{\partial\bar{\mathcal{F}}_n}{\partial t_\eta}(0,0)\right),$$

i.e., (4.3) holds.

As shown in Remark 4.9, we only need to take subsequence $\{n_j\}_{j\in\mathbb{N}}$ satisfying

$$\lim_{j\to\infty}\frac{-\left(\left\langle G_\Psi^{(n_j)}, D_\Psi^{(n_j)}\right\rangle + \left\langle G_\eta^{(n_j)}, D_\eta^{(n_j)}\right\rangle\right)}{\left|\left\langle G_\Psi^{(n_j)}, M_{\Psi^{(n_j)}}^{\eta^{(n_j)}}(G_\Psi^{(n_j)})\right\rangle\right|^a + \left|\left\langle G_\eta^{(n_j)}, M_{\eta^{(n_j)}}(G_\eta^{(n_j)})\right\rangle\right|^a} = \delta > 0$$

into account.

We observe that the corresponding $t_\Psi^{(n_j)}$ has only four options:

$$t_\Psi^{(n_j)} = \max(t_\Psi^{\text{initial}}, t_\Psi^{\min}),\ t_\Psi^{(n_j)} = \frac{\theta_\Psi^{(n_j)}}{\|D_\Psi^{(n_j)}\|},\ t_\Psi^{(n_j)} = -\frac{1}{c_{n_j,1}}\frac{\partial\mathcal{F}_{n_j}}{\partial t_\Psi}(0,0),\ t_\Psi^{(n_j)} = \frac{1}{\underline{c}}t_\eta^{(n_j)}.$$

Consequently, there exists a subsequence of $\{n_j\}_{j\in\mathbb{N}}$, which is also denoted by $\{n_j\}_{j\in\mathbb{N}}$ for convenience, such that one of the following four cases holds.

Case 1. $t_\Psi^{(n_j)} = \max(t_\Psi^{\text{initial}}, t_\Psi^{\min})$. Obviously

$$\sum_{j=0}^{\infty} t_\Psi^{(n_j)} \geq \sum_{j=0}^{\infty} t_\Psi^{\min} = +\infty,$$

which together with Remark 4.9 yields the conclusion.

Case 2. $t_\Psi^{(n_j)} = \frac{\theta_\Psi^{(n_j)}}{\|D_\Psi^{(n_j)}\|}$. If

$$\varliminf_{j\to\infty} t_\Psi^{(n_j)} > 0,$$

then Lemma 4.8 leads to the conclusion. Otherwise, there exists a subsequence of $\{n_j\}_{j\in\infty}$ also denoted by $\{n_j\}_{j\in\mathbb{N}}$ such that $\lim_{j\to\infty}\frac{\theta_\Psi^{(n_j)}}{\|D_\Psi^{(n_j)}\|} = \lim_{j\to\infty} t_\Psi^{(n_j)} = 0$.

We first prove that there holds

$$\begin{aligned}&\bar{\mathcal{F}}_n(t_\Psi, t_\eta) - \bar{\mathcal{F}}_n(0,0) - t_\Psi\frac{\partial\bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0) - t_\eta\frac{\partial\bar{\mathcal{F}}_n}{\partial t_\eta}(0,0) - \frac{1}{2}c_{n,1}t_\Psi^2 - \frac{1}{2}c_{n,2}t_\eta^2\\ &= O(t_\Psi^2\|D_\Psi^{(n)}\|^2 + t_\eta^2\|D_\eta^{(n)}\|_{sF}^2)\end{aligned} \tag{4.21}$$

when (t_Ψ, t_η) satisfies (4.9).

For convenience, we denote by $\bar{\mathcal{F}}_{n,s}(t) = \bar{\mathcal{F}}_n(t, st)$, $\Psi^{(n)}(t) = \mathrm{ortho}(\Psi^{(n)}, D_\Psi^{(n)}, t)$, $\eta^{(n)}(t) = \eta^{(n)} + tD_\eta^{(n)}$. By Assumptions 4.2 and 4.4, (4.20) and $\dot{\Psi}^{(n)}(0) = D_\Psi^{(n)}$, we have

$$
\begin{aligned}
&|\bar{\mathcal{F}}'_{n,s}(t) - \bar{\mathcal{F}}'_{n,s}(0)| \\
&= \Big| 2\,\mathrm{Re}\left\langle \mathcal{F}_\Psi(\Psi^{(n)}(t), \eta^{(n)}(st)), \dot{\Psi}^{(n)}(t) \right\rangle + s\left\langle \nabla_\eta \mathcal{F}(\Psi^{(n)}(t), \eta^{(n)}(st)), D_\eta^{(n)} \right\rangle \\
&\quad - 2\,\mathrm{Re}\left\langle \mathcal{F}_\Psi(\Psi^{(n)}(0), \eta^{(n)}(0)), \dot{\Psi}^{(n)}(0) \right\rangle - s\left\langle \nabla_\eta \mathcal{F}(\Psi^{(n)}(0), \eta^{(n)}(0)), D_\eta^{(n)} \right\rangle \Big| \\
&\le \left| 2\left\langle \mathcal{F}_\Psi(\Psi^{(n)}(t), \eta^{(n)}(st)), \dot{\Psi}^{(n)}(t) - \dot{\Psi}^{(n)}(0) \right\rangle \right| \\
&\quad + \left| 2\left\langle \mathcal{F}_\Psi(\Psi^{(n)}(t), \eta^{(n)}(st)) - \nabla_\Psi \mathcal{F}(\Psi^{(n)}(0), \eta^{(n)}(0)), \dot{\Psi}^{(n)}(0) \right\rangle \right| \\
&\quad + s\left| \left\langle \nabla_\eta \mathcal{F}(\Psi^{(n)}(t), \eta^{(n)}(st)) - \nabla_\eta \mathcal{F}(\Psi^{(n)}(0), \eta^{(n)}(0)), D_\eta^{(n)} \right\rangle \right| \\
&\le 2C_0C_2t\|D_\Psi^{(n)}\|^2 + 2L_0(C_1t\|D_\Psi^{(n)}\| + st\|D_\eta^{(n)}\|_{sF})\|D_\Psi^{(n)}\| \\
&\quad + L_0(C_1st\|D_\Psi^{(n)}\| + s^2t\|D_\eta^{(n)}\|_{sF})\|D_\eta^{(n)}\|_{sF}
\end{aligned}
$$

for any $s, t > 0$, where Proposition 2.1 and Theorem 3.2 are used in the last inequality. Applying Young inequality, we obtain that

$$
\begin{aligned}
&|\bar{\mathcal{F}}'_{n,s}(t) - \bar{\mathcal{F}}'_{n,s}(0)| \\
&\le 2C_0C_2t\|D_\Psi^{(n)}\|^2 + 2L_0\left(C_1t\|D_\Psi^{(n)}\|^2 + \frac{t\|D_\Psi^{(n)}\|^2 + s^2t\|D_\eta^{(n)}\|_{sF}^2}{2}\right) \\
&\quad + L_0\left(C_1\frac{t\|D_\Psi^{(n)}\|^2 + s^2t\|D_\eta^{(n)}\|_{sF}^2}{2} + s^2t\|D_\eta^{(n)}\|_{sF}^2\right) \\
&\le \tilde{C}t(\|D_\Psi^{(n)}\|^2 + s^2\|D_\eta^{(n)}\|_{sF}^2),
\end{aligned}
$$

where

$$
\tilde{C} = \max\left(2C_0C_2 + L_0 + \frac{5}{2}C_1L_0, 2L_0 + \frac{1}{2}C_1L_0\right).
$$

Hence we have

$$
\begin{aligned}
&\left|\bar{\mathcal{F}}_n(t, st) - \bar{\mathcal{F}}_n(0,0) - t\frac{\partial \bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0) - st\frac{\partial \bar{\mathcal{F}}_n}{\partial t_\eta}(0,0)\right| \\
&\le \int_0^t \left|\bar{\mathcal{F}}'_{n,s}(\tau) - \bar{\mathcal{F}}'_{n,s}(0)\right| \mathrm{d}\tau \\
&\le \tilde{C}t^2(\|D_\Psi^{(n)}\|^2 + s^2\|D_\eta^{(n)}\|_{sF}^2)
\end{aligned}
$$

for any $s, t > 0$. Therefore, if (t_Ψ, t_η) satisfies (4.9), then we arrive at (4.21) by (4.15).

By the definition of $(\theta_\Psi^{(n_j)}, \theta_\eta^{(n_j)})$, Assumption 4.3, (4.9), (4.14) and (4.17), for any n_j large enough, there exists

$$
t_\Psi^{*,n_j} \in \left(0, \frac{\theta_\Psi^{(n_j)}}{\|D_\Psi^{(n_j)}\|} + \frac{1}{n_j}\right),\ t_\eta^{*,n_j} = s_{n_j}^* t_\Psi^{*,n_j},\ \underline{c} \le s_{n_j}^* \le \bar{c} \tag{4.22}
$$

such that

$$
\begin{aligned}
&O((t_\Psi^{*,n_j})^2\|D_\Psi^{(n_j)}\|^2+(t_\eta^{*,n_j})^2\|D_\eta^{(n_j)}\|_{sF}^2)\\
&=\mathcal{F}_{n_j}(t_\Psi^{*,n_j},t_\eta^{*,n_j})-\mathcal{F}_{n_j}(0,0)-t_\Psi^{*,n_j}\frac{\partial\mathcal{F}_{n_j}}{\partial t_\Psi}(0,0)-t_\eta^{*,n_j}\frac{\partial\mathcal{F}_{n_j}}{\partial t_\eta}(0,0)\\
&\quad-\frac{1}{2}c_{n,1}(t_\Psi^{*,n_j})^2-\frac{1}{2}c_{n,2}(t_\eta^{*,n_j})^2\\
&>-\frac{\nu}{2}\left(t_\Psi^{*,n_j}\frac{\partial\mathcal{F}_{n_j}}{\partial t_\Psi}(0,0)+t_\eta^{*,n_j}\frac{\partial\mathcal{F}_{n_j}}{\partial t_\eta}(0,0)\right)\\
&\geq\frac{-\frac{\nu}{2}\min(1,\underline{c})\min(1,1/\bar{c})\left(\frac{\partial\mathcal{F}_{n_j}}{\partial t_\Psi}(0,0)+\frac{\partial\mathcal{F}_{n_j}}{\partial t_\eta}(0,0)\right)}{\left|\left\langle G_\Psi^{(n_j)},M_{\Psi^{(n_j)}}^{\eta^{(n_j)}}(G_\Psi^{(n_j)})\right\rangle\right|^a+\left|\left\langle G_\eta^{(n_j)},M_{\eta^{(n_j)}}(G_\eta^{(n_j)})\right\rangle\right|^a}\\
&\quad\cdot\frac{1}{\left(\|D_\Psi^{(n_j)}\|^2+\|D_\eta^{(n_j)}\|_{sF}^2\right)^{1/2}}\left((t_\Psi^{*,n_j})^2\|D_\Psi^{(n_j)}\|^2+(t_\eta^{*,n_j})^2\|D_\eta^{(n_j)}\|_{sF}^2\right)^{1/2}\\
&\quad\cdot\left(\left|\left\langle G_\Psi^{(n_j)},M_{\Psi^{(n_j)}}^{\eta^{(n_j)}}(G_\Psi^{(n_j)})\right\rangle\right|^a+\left|\left\langle G_\eta^{(n_j)},M_{\eta^{(n_j)}}(G_\eta^{(n_j)})\right\rangle\right|^a\right)\\
&\geq\frac{1}{4C}\nu\delta\min(1,\underline{c})\min(1,1/\bar{c})\min(a_\Psi^a,a_\eta^a)\left((t_\Psi^{*,n_j})^2\|D_\Psi^{(n_j)}\|^2+(t_\eta^{*,n_j})^2\|D_\eta^{(n_j)}\|_{sF}^2\right)^{1/2}\\
&\quad\cdot\left(\|G_\Psi^{(n_j)}\|^{2a}+\|G_\eta^{(n_j)}\|_{sF}^{2a}\right),
\end{aligned}
$$

i.e.,

$$
\begin{aligned}
&O(((t_\Psi^{*,n_j})^2\|D_\Psi^{(n_j)}\|^2+(t_\eta^{*,n_j})^2\|D_\eta^{(n_j)}\|_{sF}^2)^{1/2})\\
&\geq\frac{\nu\delta\min(1,\underline{c})\min(1,1/\bar{c})\min(a_\Psi^a,a_\eta^a)}{4C}\cdot\left(\|G_\Psi^{(n_j)}\|^{2a}+\|G_\eta^{(n_j)}\|_{sF}^{2a}\right).
\end{aligned}
\tag{4.23}
$$

We see from $\lim_{j\to\infty}\frac{\theta_\Psi^{(n_j)}}{\|D_\Psi^{(n_j)}\|}=0$, (4.17) and (4.22), that

$$
\lim_{j\to\infty}\left((t_\Psi^{*,n_j})^2\|D_\Psi^{(n_j)}\|^2+(t_\eta^{*,n_j})^2\|D_\eta^{(n_j)}\|_{sF}^2\right)^{1/2}=0.
$$

Let $j\to\infty$ in (4.23), we get

$$
0\geq\frac{\nu\delta\min(1,\underline{c})\min(1,1/\bar{c})\min(a_\Psi^a,a_\eta^a)}{4C}\lim_{j\to\infty}\left(\left(\|G_\Psi^{(n_j)}\|^{2a}+\|G_\eta^{(n_j)}\|_{sF}^{2a}\right)\right),
$$

which produces the conclusion.

Case 3. $t_\Psi^{(n_j)}=-\frac{1}{c_{n_j,1}}\frac{\partial\mathcal{F}_{n_j}}{\partial t_\Psi}(0,0)=-\frac{\left\langle G_\Psi^{(n_j)},D_\Psi^{(n_j)}\right\rangle}{c_{n_j,1}}$. We get from Assumption 4.6 that $t_\eta^{(n_j)}$ has only three options:

$$
t_\eta^{(n_j)}=\max(t_\eta^{\text{initial}},t_\eta^{\min}),\ t_\eta^{(n_j)}=\frac{\theta_\eta^{(n_j)}}{\|D_\eta^{(n_j)}\|},\ t_\eta^{(n_j)}=-\frac{1}{c_{n_j,2}}\frac{\partial\mathcal{F}_{n_j}}{\partial t_\eta}(0,0).
$$

If $t_\eta^{(n_j)}$ is one of the first two options, the similar arguments in Cases 1 and 2 can be applied to $t_\eta^{(n_j)}$. Thus, let $t_\eta^{(n_j)}=-\frac{1}{c_{n_j,2}}\frac{\partial\mathcal{F}_{n_j}}{\partial t_\eta}(0,0)=-\frac{\left\langle G_\eta^{(n_j)},D_\eta^{(n_j)}\right\rangle}{c_{n_j,2}}$. Then we obtain

from Assumptions 4.5 and 4.7, (4.9), and (4.14) that

$$\begin{aligned}
&(1+\bar{c})t_{\Psi}^{(n_j)} \\
&\geq t_{\Psi}^{(n_j)} + t_{\eta}^{(n_j)} \\
&\geq -\frac{\left\langle G_{\Psi}^{(n_j)}, D_{\Psi}^{(n_j)} \right\rangle + \left\langle G_{\eta}^{(n_j)}, D_{\eta}^{(n_j)} \right\rangle}{\bar{C}(\|D_{\Psi}^{(n_j)}\|^2 + \|D_{\eta}^{(n_j)}\|_{sF}^2)} \\
&= \frac{-\left(\left\langle G_{\Psi}^{(n_j)}, D_{\Psi}^{(n_j)} \right\rangle + \left\langle G_{\eta}^{(n_j)}, D_{\eta}^{(n_j)} \right\rangle\right)}{\left|\left\langle G_{\Psi}^{(n_j)}, M_{\Psi^{(n_j)}}^{\eta^{(n_j)}}(G_{\Psi}^{(n_j)}) \right\rangle\right|^a + \left|\left\langle G_{\eta}^{(n_j)}, M_{\eta^{(n_j)}}(G_{\eta}^{(n_j)}) \right\rangle\right|^a} \\
&\quad \cdot \frac{\left|\left\langle G_{\Psi}^{(n_j)}, M_{\Psi^{(n_j)}}^{\eta^{(n_j)}}(G_{\Psi}^{(n_j)}) \right\rangle\right|^a + \left|\left\langle G_{\eta}^{(n_j)}, M_{\eta^{(n_j)}}(G_{\eta}^{(n_j)}) \right\rangle\right|^a}{\bar{C}(\|D_{\Psi}^{(n_j)}\|^2 + \|D_{\eta}^{(n_j)}\|_{sF}^2)} \\
&\geq \frac{\delta \min(a_{\Psi}^a, a_{\eta}^a)}{2\bar{C}C^2} \left(\|G_{\Psi}^{(n_j)}\|^{2a} + \|G_{\eta}^{(n_j)}\|_{sF}^{2a}\right)
\end{aligned}$$

provided $j \gg 1$. Consequently, either

$$\sum_{j=0}^{\infty} t_{\Psi}^{(n_j)} = \infty$$

or

$$\underline{\lim}_{j \to +\infty} (\|\nabla_{\Psi} \mathcal{F}(\Psi^{(n_j)}, \eta^{(n_j)})\|^{2a} + \|\nabla_{\eta} \mathcal{F}(\Psi^{(n_j)}, \eta^{(n_j)})\|_{sF}^{2a}) = 0,$$

which leads to the conclusion.

Case 4. $t_{\eta}^{(n_j)} = \frac{1}{\underline{c}} t_{\eta}^{(n_j)}$. We observe that the corresponding $t_{\Psi}^{(n_j)}$ has only two options:

$$t_{\eta}^{(n_j)} = \max(t_{\eta}^{\text{initial}}, t_{\eta}^{\min}), \ t_{\eta}^{(n_j)} = \frac{\theta_{\eta}^{(n_j)}}{\|D_{\eta}^{(n_j)}\|_{sF}}.$$

Thus applying similar arguments in Cases 1 and 2 for $t_{\Psi}^{(n_j)}$ to $t_{\eta}^{(n_j)}$, we complete the proof. ∎

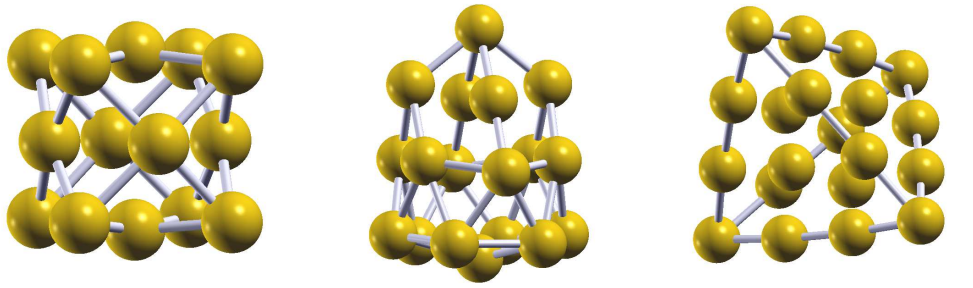
5. Numerical experiments. In this section, we apply the PCG method and its restarted versions to simulate several gold clusters (see Figure 1 for their configurations) and two complicated multicomponent periodic systems (see Figure 2 for their configurations). We implement the PCG method and its restarted versions in the software package Quantum ESPRESSO [33]. All calculations are carried out on LSSC-IV in the State Key Laboratory of Scientific and Engineering Computing of the Chinese Academy of Sciences.

In our numerical experiments, we do not restrict the step sizes to satisfy (4.9) for some given parameters $\underline{c}$ and $\bar{c}$, which can be viewed as $\underline{c} = 0$, $\bar{c} = +\infty$. Although (4.9) is necessary in our theoretical analysis, numerical results show that the step sizes can be more relaxed. Therefore, we directly apply the step size strategies (S1), (S2) or (S3) to get the step sizes in the numerical simulations.

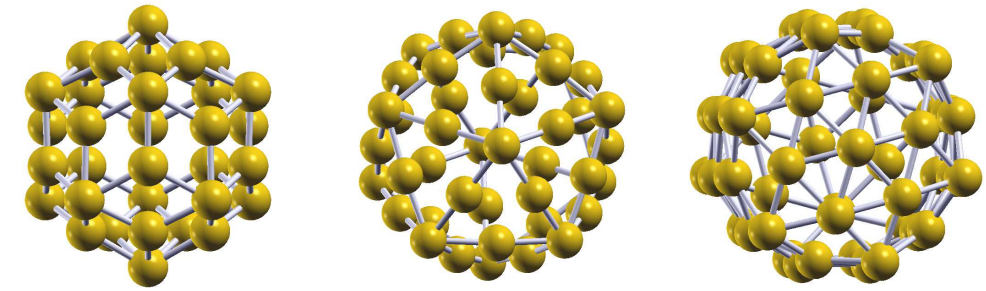
In the following tables and figures, PCG-S1, PCG-S2 and PCG-S3 stand for the corresponding PCG method (Algorithm 4.2) when the step size strategy (S1), strategy (S2) and strategy (S3) are applied, respectively. We denote the restarted

Fig. 1: The configurations of the gold clusters

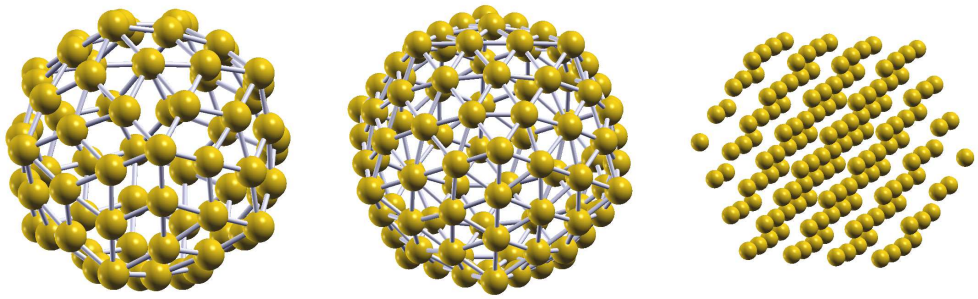
(a) Au_{14} (b) Au_{18} (c) Au_{20}



(d) Au_{32} (e) Au_{42} (f) Au_{50}



(g) Au_{72} (h) Au_{92} (i) Au_{147}



(j) Au_{309}

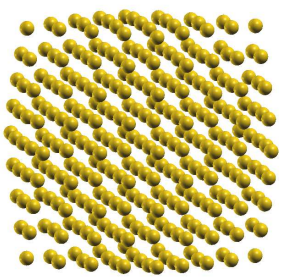
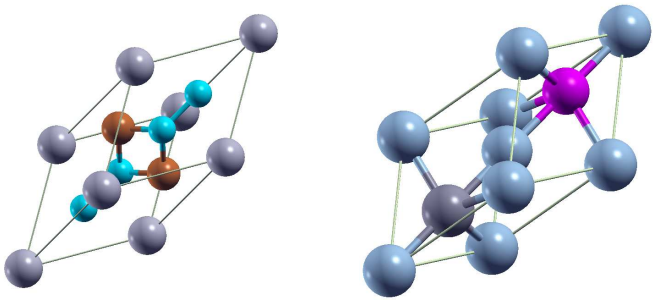


Fig. 2: The configurations in the unit cell for the multicomponent periodic systems

(a) $NdCu_2Si_2$ (b) AlCrTiV



versions Algorithm 4.3 and Algorithm 4.4 by PCG-S$\star$-r1 and PCG-S$\star$-r2 respectively, where $\star$ can be 1, 2 or 3. We mention that "Error" for the SCF iteraions is the error of density and "Error" for the PCG methods is $\left(\|\frac{1}{2}\nabla_\Psi \mathcal{F}\|^2 + \|\nabla_\eta \mathcal{F}\|_{sF}^2\right)^{1/2}$.

We will compare our PCG methods with the SCF iterations. It is known that we have to solve a linear eigenvalue problem at each SCF iteration, for which the Davidson iterative diagonalization and the CG diagonalization are commonly used in Quantum ESPRESSO. The Davidson iterative diagonalization is faster, but the CG diagonalization uses less memory and is more robust [33].

We list all the parameters used in our numerical experiments. The Ultrasoft pseudopotentials and the Gaussian smearing with $\sigma = 0.05$ Ry are applied for gold clusters. We use the DY approach to get the CG parameter and the QR strategy as (4.11) for the orthogonalization operation. We apply $\theta^{(n)} = \min\{0.8, \sqrt{\|D_\Psi^{(n)}\|_\infty^2 + \|D_\eta^{(n)}\|_{sF,\infty}^2}\}$ for strategies (S1) and (S2) and $\theta_\Psi^{(n)} = \min\{0.8, \|D_\Psi^{(n)}\|_\infty\}$, $\theta_\eta^{(n)} = \min\{0.8, \|D_\eta^{(n)}\|_{sF,\infty}\}$ for strategy (S3). We set $t^{\min} = t_\Psi^{\min} = t_\eta^{\min} = 0.001$ and initial trial step sizes $t^{\text{trial}} = t_\Psi^{\text{trial}} = t_\eta^{\text{trial}} = 0.4$. For the restarted versions, we set $\gamma = 0.5$ and $a = 1$. The convergence criterion is

$$\left(\|\frac{1}{2}\nabla_\Psi \mathcal{F}\|^2 + \|\nabla_\eta \mathcal{F}\|_{sF}^2\right)^{1/2} < 1.0 \times 10^{-5}$$

for the PCG method and its restarted versions, and the convergence threshold for density is 1.0×10^{-9} for the SCF iterations. For the SCF iterations, We apply the Broyden mixing method. The initial guess for the wavefunctions is generated by the superposition of atomic orbitals [33] if not specified.

We see that whether or not to restart has almost no effect for the simulation of gold clusters for the strategies (S2) and (S3). As a result, we mainly show the numerical results obtained by the PCG method (Algorithm 4.2) for gold clusters. In addition, we will also mention some improvement of the restarting approach for the strategy (S1) in Figure 3.

First, we take a look at the results of all the gold clusters. The results obtained by the PCG method (Algorithm 4.2) based on different step size strategies are listed in Table 1. In Table 1, "Iter." means the number of iterations required to terminate the algorithm and "A.T.P.I" is the average CPU time required per iteration. As shown in

Table 1, the strategy (S3) with different step sizes for Ψ and η is indeed the best. More precisely, the strategy (S3) needs less iteration and the CPU time to achieve similar accuracy than the strategies (S1) and (S2) with same step sizes for Ψ and η, especially for large systems. We see that the strategies (S2) and (S3) are more expensive than the strategy (S1) per iteration. However, by comparing the strategies (S1) and (S2), we see that the strategy (S1) need more iterations to achieve the same accuracy than the strategy (S2). Even the iterations for Au_{18}, Au_{72}, Au_{92}, Au_{147} and Au_{309} do not converge after 200 iterations under the strategy (S1). We point out that whether or not to restart has no effect on the strategies (S2) and (S3) for the simulation of these gold clusters under the convergence criterion discussed in this section. However, it will improve the convergence of the iteration a little for the strategy (S1). If we restart the PCG method as Algorithm 4.3, the calculations for Au_{18}, Au_{72} and Au_{92} can also converge under the strategy (S1). Due to limited space, we only show the results of Au_{92} obtained by the restarted PCG method I (Algorithm 4.3) later.

To compare the three step size strategies more clearly, we take Au_{92} as an example and show the convergence curves for $\mathcal{F} - \mathcal{F}_{\min}$, $\frac{1}{2}\|\nabla_\Psi \mathcal{F}\|$ and $\|\nabla_\eta \mathcal{F}\|_{sF}$ in Figure 3, where $\mathcal{F}_{\min}$ is a high-accuracy approximation of the exact total energy. We also illustrate the benefit of the restarting approach for the strategy (S1). First, the strategy (S3) is indeed faster than the other two strategies. Secondly, by comparing the convergence curves for the error of the energy, we see that the strategy (S1) is not much different from the strategy (S2), and the strategy (S1) seems to be better when the energy has not converged. But there may be some fluctuation for the strategy (S1) when the energy almost converges. From the convergence curves for $\frac{1}{2}\|\nabla_\Psi \mathcal{F}\|$ and $\|\nabla_\eta \mathcal{F}\|_{sF}$, we see that the descent speed of the gradient obtained by the strategy (S1) slows down suddenly when the energy almost converges and then is much smaller than the strategy (S2). Finally, by comparing PCG-S1 and PCG-S1-r1, we find that the restarting approach does improve the convergence of the iteration for the strategy (S1).

We conclude from the above that the strategy (S3) seems to be the best one among the three strategies. We then choose the PCG method based on the step size strategy (S3) to be compared with the SCF iterations based on the CG diagonalization. The detailed results are shown in Table 2. We see from Table 2 that, apart from Au_{14}, the PCG method converges faster than the SCF iterations, especially for large scale systems. For instance, the PCG method converges in half the CPU time of SCF iterations for Au_{42}, and the PCG method converges in less than 1/3 the CPU time of the SCF iterations for Au_{147}. We also mention that the energy obtained by the PCG method is slightly smaller than that obtained by SCF iterations for Au_{20}, Au_{42}, Au_{50}, Au_{92} and Au_{309}, which means that SCF iterations may require a smaller convergence threshold to obtain the same energy obtained by the PCG method. However, SCF iterations has already cost more CPU time even with the accuracy in the table.

Now, we show the numerical results for the two complicated periodic systems shown in Figure 2. Different from the gold clusters, for these two systems, the spin polarization is taken into account and the cases using different initial guesses of wavefunctions are tested. Since these two systems show more obvious metallicity, more smearing strategies may be used. Here, we consider the Gaussian smearing and the Marzari–Vanderbilt smearing, which are some typical smearing functions used in the simulation of metallic systems. The detailed results are reported in Tables 3 and 4. Here, $N_k = 2|\mathcal{K}|$, "atomic" means that the initial guess of wavefunctions is generated by the superposition of atomic orbitals, and "atomic+random" means that the initial guess of wavefunctions is generated by the superposition of atomic orbitals plus a

Table 1: The numerical results for gold clusters obtained by the PCG method (Algorithm 4.2) based on different step size strategies.

Algorithm	Energy (Ry)	Iter.	Error	CPU time (s)	A.T.P.I (s)
Au_{14} $N_G = 322453$ $N = 92$ $cores = 36$					
PCG-S1	-1194.49861028	90	9.3E-6	587.0	6.52
PCG-S2	-1194.49861028	50	9.7E-6	397.7	7.95
PCG-S3	-1194.49861028	37	8.5E-6	299.2	8.09
Au_{18} $N_G = 322453$ $N = 119$ $cores = 36$					
PCG-S1	-1536.01945578	200	1.4E-5	1626.6	8.13
PCG-S2	-1536.01945578	62	7.2E-6	647.3	10.44
PCG-S3	-1536.01945578	37	8.9E-6	384.6	10.39
Au_{20} $N_G = 322453$ $N = 132$ $cores = 36$					
PCG-S1	-1706.76524000	109	9.1E-6	963.3	8.84
PCG-S2	-1706.76524000	55	8.3E-6	621.6	11.30
PCG-S3	-1706.76524000	38	9.1E-6	429.7	11.31
Au_{32} $N_G = 429409$ $N = 211$ $cores = 36$					
PCG-S1	-2731.11762824	90	8.3E-6	1808.6	20.10
PCG-S2	-2731.11762824	48	8.4E-6	1270.7	26.47
PCG-S3	-2731.11762824	38	8.3E-6	1019.1	26.82
Au_{42} $N_G = 429409$ $N = 277$ $cores = 36$					
PCG-S1	-3584.66580292	78	1.0E-6	2133.8	27.36
PCG-S2	-3584.66580292	55	5.8E-6	2011.6	36.57
PCG-S3	-3584.66580292	39	8.5E-6	1390.6	35.66
Au_{50} $N_G = 429409$ $N = 330$ $cores = 36$					
PCG-S1	-4267.69535810	114	6.8E-6	3700.7	32.46
PCG-S2	-4267.69535810	58	9.4E-6	2626.5	45.28
PCG-S3	-4267.69535810	39	9.2E-6	1786.2	45.80
Au_{72} $N_G = 556667$ $N = 475$ $cores = 36$					
PCG-S1	-6145.78233806	200	1.1E-4	13959.2	69.80
PCG-S2	-6145.78233806	89	9.8E-6	8557.1	96.15
PCG-S3	-6145.78233806	40	9.0E-6	3760.7	94.02
Au_{92} $N_G = 556667$ $N = 607$ $cores = 36$					
PCG-S1	-7853.07110320	200	2.3E-5	19697.0	98.49
PCG-S2	-7853.07110320	91	7.9E-6	12229.9	134.39
PCG-S3	-7853.07110320	40	9.8E-6	5535.4	138.39
Au_{147} $N_G = 1320073$ $N = 971$ $cores = 72$					
PCG-S1	-12547.62980551	200	4.4E-5	37056.7	185.28
PCG-S2	-12547.62980551	88	9.7E-6	23166.6	263.26
PCG-S3	-12547.62980551	42	9.0E-6	11193.5	266.51
Au_{309} $N_G = 1320073$ $N = 2040$ $cores = 72$					
PCG-S1	-26379.41930504	200	1.5E-4	134638.7	673.19
PCG-S2	-26379.41930504	124	9.1E-6	119025.8	959.89
PCG-S3	-26379.41930507	51	6.8E-6	49831.2	977.08

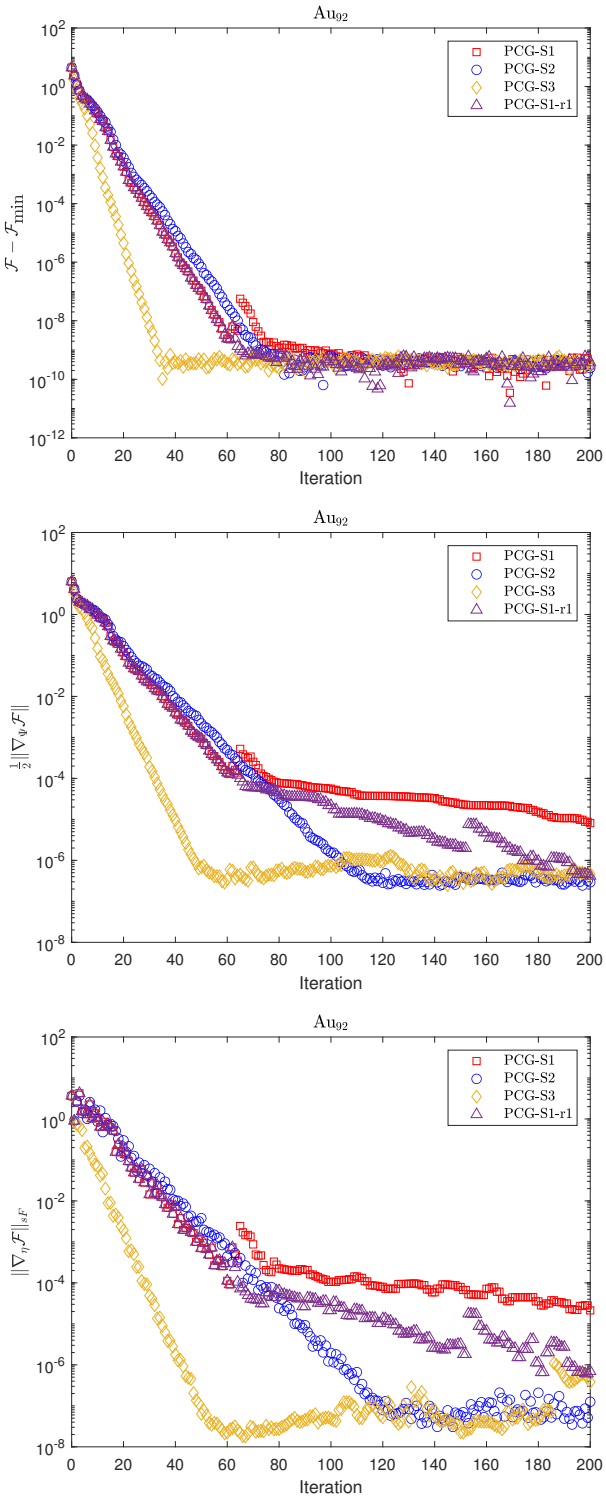


Fig. 3: Convergence curves for $\mathcal{F}-\mathcal{F}_{\min}$, $\frac{1}{2}\|\nabla_\Psi \mathcal{F}\|$ and $\|\nabla_\eta \mathcal{F}\|_{sF}$ obtained by different step size strategies for Au_{92}.

Table 2: Comparison of the SCF iterations based on the CG diagonalization and the PCG method based on the step size strategy (S3). The density mixing factor for the SCF iterations is 0.3.

Algorithm	Energy (Ry)	Iter.	Error	CPU time (s)
Au_{14}	$N_G = 322453$	$N = 92$	$cores = 36$	
SCF	-1194.49861028	16	9.5E-10	271.9
PCG-S3	-1194.49861028	37	8.5E-6	299.2
Au_{18}	$N_G = 322453$	$N = 119$	$cores = 36$	
SCF	-1536.01945578	18	5.5E-10	452.7
PCG-S3	-1536.01945578	37	8.9E-6	384.6
Au_{20}	$N_G = 322453$	$N = 132$	$cores = 36$	
SCF	-1706.76523999	15	8.3E-10	470.2
PCG-S3	-1706.76524000	38	9.1E-6	429.7
Au_{32}	$N_G = 429409$	$N = 211$	$cores = 36$	
SCF	-2731.11762824	16	3.5E-10	1476.1
PCG-S3	-2731.11762824	38	8.3E-6	1019.1
Au_{42}	$N_G = 429409$	$N = 277$	$cores = 36$	
SCF	-3584.66580291	20	2.2E-11	2870.1
PCG-S3	-3584.66580292	39	8.5E-6	1390.6
Au_{50}	$N_G = 429409$	$N = 330$	$cores = 36$	
SCF	-4267.69535809	17	7.8E-10	3629.5
PCG-S3	-4267.69535810	39	9.2E-6	1786.2
Au_{72}	$N_G = 556667$	$N = 475$	$cores = 36$	
SCF	-6145.78233806	24	1.9E-10	10766.7
PCG-S3	-6145.78233806	40	9.0E-6	3760.7
Au_{92}	$N_G = 556667$	$N = 607$	$cores = 36$	
SCF	-7853.07110315	21	4.8E-10	16142.4
PCG-S3	-7853.07110320	40	9.8E-6	5535.4
Au_{147}	$N_G = 1320073$	$N = 971$	$cores = 72$	
SCF	-12547.62980551	30	3.8E-10	39669.2
PCG-S3	-12547.62980551	42	9.0E-6	11193.5
Au_{309}	$N_G = 1320073$	$N = 2040$	$cores = 72$	
SCF	-26379.41930501	23	3.5E-10	154451.0
PCG-S3	-26379.41930507	51	6.8E-6	49831.2

superimposed "randomization" of atomic orbitals [33]. We observe from Tables 3 and 4 that, for both the two smearing methods, except for the system $NdCu_2Si_2$ with the initial guess of wavefunctions being given by the superposition of atomic orbitals, the SCF iterations fail to converge after 500 iterations. We also see that both the PCG method and the restarded PCG methods can obtain convergent approximations for both the two systems, no matter what kind of initial guesses and smearing methods are used. Comparing the results for PCG-S3 with the results for PCG-S3-r1 and PCG-S3-r2, we observe that the restarting strategy does accelerate the convergence of the PCG method except for the system $NdCu_2Si_2$ calculated by PCG-S3-r2 with the initial guesses of wavefunctions being given by "atomic+random" and the Gaussian smearing. Comparing the results for PCG-S3-r1 with the results for PCG-S3-r2, we

Table 3: Comparison of the SCF iterations based on the Davidson iterative diagonalization, the PCG method and the restarted PCG methods. The density mixing factor for the SCF iterations is 0.4, and the Gaussian smearing with $\sigma = 0.01$ Ry is applied.

Algorithm	Initial orbitals	Energy (Ry)	Iter.	Error
$NdCu_2Si_2$	$N_G = 3837$ $N = 36$ $N_k = 576$ $cores = 36$			
SCF	atomic	-1368.00296219	24	1.9E-10
	atomic+random	-1367.99467076	500	6.8E-6
PCG-S3	atomic	-1368.00296213	440	9.9E-6
	atomic+random	-1367.99713429	255	9.8E-6
PCG-S3-r1	atomic	-1368.00296213	313	9.9E-6
	atomic+random	-1367.99713429	239	9.5E-6
PCG-S3-r2	atomic	-1368.00296205	308	8.0E-6
	atomic+random	-1367.99713429	290	8.8E-6
AlCrTiV	$N_G = 1759$ $N = 25$ $N_k = 144$ $cores = 36$			
SCF	atomic	-479.31372455	500	2.1E-4
	atomic+random	-479.31491981	500	7.0E-4
PCG-S3	atomic	-479.36755754	200	9.9E-6
	atomic+random	-479.36755753	283	9.3E-6
PCG-S3-r1	atomic	-479.36755753	136	9.2E-6
	atomic+random	-479.36755753	234	9.6E-6
PCG-S3-r2	atomic	-479.36755753	109	8.5E-6
	atomic+random	-479.36755754	115	8.0E-6

also see that the second restarting approach (Algorithm 4.4) is better than the first restarting approach (Algorithm 4.3) for the system AlCrTiV, but the first restarting approach is better than the second restarting approach for the system $NdCu_2Si_2$. We conclude that the PCG method and the restarted PCG methods are more stable when different initial orbitals are used and our methods are suitable for different smearing functions.

6. Concluding remarks. In this paper, we have first investigated the energy minimization model of the ensemble Kohn-Sham density functional theory from a mathematical aspect, in which the pseudo-eigenvalue matrix and the general smearing approach are involved. We have shown the invariance and the existence of the minimizer of the energy functional and proposed a preconditioned conjugate gradient method to solve the numerical approximations of the energy minimization problem. In particular, we have presented an adaptive double step size strategy since the iterative behavior for Ψ and η may be different. Under some mild and reasonable assumptions, we have obtained the global convergence of the PCG algorithm based on the adaptive double step size strategy. We have reported a large number of numerical experiments which can not only verify our theory, but also show the superiority over the traditional SCF iterations. In particular, our numerical experiments have demonstrated that our algorithm can produce convergent numerical approximations for some metallic systems, for which the traditional self-consistent field iterations fails to converge.

Appendix A. Gradient of the energy functional. In this appendix, we introduce the gradient of $\mathcal{F}$ with respect to Ψ and η. Assume that the exchange-

Table 4: Comparison of the SCF iterations based on the Davidson iterative diagonalization, the PCG method and the restarted PCG methods. The density mixing factor for the SCF iterations is 0.4, and the Marzari–Vanderbilt smearing with $\sigma = 0.01$ Ry is applied.

Algorithm	Initial orbitals	Energy (Ry)	Iter.	Error
NdCu2Si2	$N_G = 3837$ $N = 36$ $N_k = 576$		$cores = 36$	
SCF	atomic	-1368.00214304	24	4.0E-10
	atomic+random	-1367.99221653	500	5.6E-6
PCG-S3	atomic	-1368.00214302	313	9.7E-6
	atomic+random	-1368.00214304	541	9.9E-6
PCG-S3-r1	atomic	-1367.99610068	309	8.0E-6
	atomic+random	-1368.00214304	226	9.4E-6
PCG-S3-r2	atomic	-1368.00214301	310	9.5E-6
	atomic+random	-1368.00214303	315	9.1E-6
AlCrTiV	$N_G = 1759$ $N = 25$ $N_k = 144$		$cores = 36$	
SCF	atomic	-479.31161107	500	9.6E-4
	atomic+random	-479.30711971	500	2.1E-3
PCG-S3	atomic	-479.36717223	204	9.8E-6
	atomic+random	-479.36717223	154	9.2E-6
PCG-S3-r1	atomic	-479.36717223	104	9.8E-6
	atomic+random	-479.36717223	128	9.2E-6
PCG-S3-r2	atomic	-479.36717223	90	8.5E-6
	atomic+random	-479.36717223	109	9.5E-6

correction functional $\mathcal{E}_{\rm xc}$ is differentiable.

Since $\Psi_{\rm k}$ is complex valued and $\mathcal{F}$ is real valued, $\mathcal{F}$ is not differentiable with respect to $\Psi_{\rm k}$. Let $\Psi_{\rm k} = \Psi_{\rm k,Re} + {\rm i}\Psi_{\rm k,Im}$, where $\Psi_{\rm k,Re}$ and $\Psi_{\rm k,Im}$ are real valued. We see that $\mathcal{F}$ is differentiable with respect to $\Psi_{\rm k,Re}$ and $\Psi_{\rm k,Im}$. Thus we apply the Wirtinger derivatives. More precisely, we view $\Psi_{\rm k}$ and $\bar{\Psi}_{\rm k}$ as two independent variables for all ${\rm k} \in \mathcal{K}$, then the energy functional (2.5) is a differentiable functional of Ψ, $\bar{\Psi}$ and η, which is still denoted by $\mathcal{F}$ for convenience, namely, $\mathcal{F}(\Psi, \bar{\Psi}, \eta)$. A direct calculation shows

$$\mathcal{F}_{\Psi_{\rm k}} = \frac{1}{2}(\mathcal{F}_{\Psi_{\rm k,Re}} - {\rm i}\mathcal{F}_{\Psi_{\rm k,Im}}).$$

We refer to [23] for more details. We use the convenient notation $\mathcal{F}(\Psi, \eta) = \mathcal{F}(\Psi, \bar{\Psi}, \eta)$ and $\mathcal{L}(\Psi, \eta, \Lambda) = \mathcal{L}(\Psi, \bar{\Psi}, \eta, \Lambda)$. Then there holds

$$\mathcal{F}_{\Psi_{\rm k}}(\Psi, \eta) = w_{\rm k} H_{\rm k}(\rho_{\Psi,\eta})\Psi_{\rm k} F_{\eta_{\rm k}}$$

and

$$\mathcal{L}_{\Psi_{\rm k}}(\Psi, \eta, \Lambda) = w_{\rm k}(H_{\rm k}(\rho_{\Psi,\eta})\Psi_{\rm k} F_{\eta_{\rm k}} - \mathcal{B}\Psi_{\rm k}\Lambda_{\rm k}),$$

where

$$H_{\rm k}(\rho) = -\frac{1}{2}({\rm ik} + \nabla)^2 + \tilde{V}_{\rm loc}(\rho) + \tilde{V}_{\rm nl}(\rho)$$

with $\tilde{V}_{\rm loc}(\rho) = V_{\rm loc} + \displaystyle\int_\Omega \frac{\rho(r)}{|\cdot - r|}{\rm d}r + V_{\rm xc}(\rho)$, $\tilde{V}_{\rm nl}(\rho) : \Psi_{\rm k} \mapsto V_{\rm nl}(\Psi_{\rm k}) + M\tilde{D}\,\langle M^*\Psi_{\rm k}\rangle$,

$V_{\rm xc}(\rho) = \dfrac{\delta \mathcal{E}_{\rm xc}}{\delta \rho}$, and

$$\tilde{D} = \int_\Omega \tilde{V}_{\rm loc}(\rho)(r)\mathcal{Q}(r)\,{\rm d}r.$$

It is clear that at any minimizer (Ψ, η), we have

$$\Lambda_{\rm k} = \langle \Psi_{\rm k}^* H(\rho_{\Psi,\eta})\Psi_{\rm k}\rangle F_{\eta_{\rm k}}.$$

Hence we set

$$\nabla_{\Psi_{\rm k}}\mathcal{F}(\Psi,\eta) = 2\mathcal{L}_{\Psi_{\rm k}}(\Psi,\eta,(\langle\Psi_{\rm k}^* H(\rho_{\Psi,\eta})\Psi_{\rm k}\rangle F_{\eta_{\rm k}})_{{\rm k}\in\mathcal{K}})$$

and $\nabla_\Psi \mathcal{F} = (\nabla_{\Psi_{\rm k}}\mathcal{F})_{{\rm k}\in\mathcal{K}}$. Obviously, $\mathcal{L}_{\Psi_{\rm k,Re}} = \mathcal{L}_{\Psi_{\rm k,Im}} = 0$ if and only if $\mathcal{L}_{\Psi_{\rm k}} = 0$.

Then we calculate $\mathcal{F}_{\eta_{\rm k}} = \left(\dfrac{\partial \mathcal{F}}{\partial \eta_{{\rm k}ij}}\right)_{i,j=1}^N$ by referring to Appendix E in [21]. We see that

$$\mathrm{d}\epsilon_{{\rm k}i} = (P_{\rm k}^* \mathrm{d}\eta_{\rm k} P_{\rm k})_{ii}, \quad i = 1,\dots,N \tag{A.1}$$

and

$$\begin{aligned}
(\mathrm{d}F_{\eta_{\rm k}})_{ij} &= \sum_{i',j'=1}^N P_{{\rm k}ii'}\left(P_{\rm k}^* f\left(\frac{\eta_{\rm k}-\mu I}{\sigma}\right)P_{\rm k}\right)_{i'j'} P^*_{{\rm k}j'j} \\
&= \sum_{i'=1}^N P_{{\rm k}ii'}P^*_{{\rm k}i'j}\frac{1}{\sigma}f'\left(\frac{\epsilon_{{\rm k}i'}-\mu}{\sigma}\right)(\mathrm{d}\epsilon_{{\rm k}i'} - \mathrm{d}\mu) \\
&\quad + \sum_{i'\neq j'} P_{{\rm k}ii'}P^*_{{\rm k}j'j}\frac{f_{{\rm k}j'} - f_{{\rm k}i'}}{\epsilon_{{\rm k}j'} - \epsilon_{{\rm k}i'}}(P_{\rm k}^*\mathrm{d}\eta_{\rm k}P_{\rm k})_{i'j'},
\end{aligned} \tag{A.2}$$

where $P = (P_{\rm k})_{{\rm k}\in\mathcal{K}} \in \left(\mathcal{O}_{\mathbb{C}}^{N\times N}\right)^{|\mathcal{K}|}$, $P_{\rm k}^*\eta_{\rm k}P_{\rm k} = {\rm Diag}(\epsilon_{\rm k1},\dots,\epsilon_{{\rm k}N})$, $f_{{\rm k}i} = f((\epsilon_{{\rm k}i}-\mu)/\sigma)$. We get from $\sum\limits_{{\rm k}\in\mathcal{K}} w_{\rm k}\,{\rm tr}\,F_{\eta_k} = N_e$ that

$$\mathrm{d}\mu = \frac{\sum_{{\rm k}\in\mathcal{K}} w_{\rm k}\sum_{i=1}^N f'\left(\frac{\epsilon_{{\rm k}i}-\mu}{\sigma}\right)\mathrm{d}\epsilon_{{\rm k}i}}{\sum_{{\rm k}\in\mathcal{K}} w_{\rm k}\sum_{i=1}^N f'\left(\frac{\epsilon_{{\rm k}i}-\mu}{\sigma}\right)}. \tag{A.3}$$

Moreover, we have

$$\begin{aligned}
\mathrm{d}\left(\sigma\,{\rm tr}\,S\left(\frac{1}{\sigma}(\eta_{\rm k}-\mu I)\right)\right) &= \sigma\sum_{i'=1}^N \mathrm{d}S\left(\frac{\epsilon_{{\rm k}i'}-\mu}{\sigma}\right) \\
&= \sum_{i'=1}^N S'\left(\frac{\epsilon_{{\rm k}i'}-\mu}{\sigma}\right)(\mathrm{d}\epsilon_{{\rm k}i'} - \mathrm{d}\mu) \\
&= \sum_{i'=1}^N \frac{1}{\sigma}(\epsilon_{{\rm k}i'}-\mu)f'\left(\frac{\epsilon_{{\rm k}i'}-\mu}{\sigma}\right)(\mathrm{d}\epsilon_{{\rm k}i'} - \mathrm{d}\mu)\,.
\end{aligned} \tag{A.4}$$

It follows from (A.2) and (A.4) that

$$
\begin{aligned}
&\frac{\partial \mathcal{F}}{\partial \eta_{\mathrm{k}ij}} \\
&= \frac{\partial \mathcal{E}}{\partial \eta_{\mathrm{k}ij}} - \sum_{\mathrm{k}'\in\mathcal{K}} w_{\mathrm{k}'}\sigma \frac{\partial \operatorname{tr} S\left(\frac{1}{\sigma}(\eta_{\mathrm{k}'}-\mu I)\right)}{\partial \eta_{\mathrm{k}ij}} \\
&= \sum_{\mathrm{k}'\in\mathcal{K}} \sum_{i',j'=1}^{N} \frac{\partial \mathcal{E}}{\partial (F_{\eta_{\mathrm{k}'}})_{i'j'}} \frac{\partial (F_{\eta_{\mathrm{k}'}})_{i'j'}}{\partial \eta_{\mathrm{k}ij}} - \sum_{\mathrm{k}'\in\mathcal{K}} w_{\mathrm{k}'}\sigma \frac{\partial \operatorname{tr} S\left(\frac{1}{\sigma}(\eta_{\mathrm{k}'}-\mu I)\right)}{\partial \eta_{\mathrm{k}ij}} \\
&= \sum_{\mathrm{k}'\in\mathcal{K}} w_{\mathrm{k}'} \sum_{i''=1}^{N} \left(\sum_{i',j'=1}^{N} \langle \psi_{\mathrm{k}'j'}, H_{\mathrm{k}'}(\rho_{\Psi,\eta})\psi_{\mathrm{k}'i'}\rangle P_{\mathrm{k}'i'i''} P^*_{\mathrm{k}'i''j'} - \epsilon_{\mathrm{k}'i''} + \mu \right) \frac{1}{\sigma} f'\left(\frac{\epsilon_{\mathrm{k}'i''}-\mu}{\sigma}\right) \frac{\partial \epsilon_{\mathrm{k}'i''}}{\partial \eta_{\mathrm{k}ij}} \\
&\quad - \frac{\partial \mu}{\partial \eta_{\mathrm{k}ij}} \sum_{\mathrm{k}\in\mathcal{K}} w_{\mathrm{k}'} \sum_{i''=1}^{N} \left(\sum_{i',j'=1}^{N} \langle \psi_{\mathrm{k}'j'}, H_{\mathrm{k}'}(\rho_{\Psi,\eta})\psi_{\mathrm{k}i'}\rangle P_{\mathrm{k}'i'i''} P^*_{\mathrm{k}'i''j'} - \epsilon_{\mathrm{k}'i''} + \mu \right) \frac{1}{\sigma} f'\left(\frac{\epsilon_{\mathrm{k}i''}-\mu}{\sigma}\right) \\
&\quad + w_{\mathrm{k}} \sum_{i''\neq j''} \left(\sum_{i',j'=1}^{N} \langle \psi_{\mathrm{k}j'}, H_{\mathrm{k}}(\rho_{\Psi,\eta})\psi_{\mathrm{k}i'}\rangle P_{\mathrm{k}i'i''} P^*_{\mathrm{k}j''j'} \right) \frac{f_{\mathrm{k}j''} - f_{\mathrm{k}i''}}{\epsilon_{\mathrm{k}j''} - \epsilon_{\mathrm{k}i''}} P^*_{\mathrm{k}i''i} P_{\mathrm{k}jj''},
\end{aligned}
$$

which together with (A.1) and (A.3) leads to

$$
\begin{aligned}
&\frac{\partial \mathcal{F}}{\partial \eta_{\mathrm{k}ij}} \\
&= w_{\mathrm{k}} \sum_{i'=1}^{N} (\langle \tilde{\psi}_{\mathrm{k}i'}, H_{\mathrm{k}}(\rho_{\tilde{\Psi},\eta_{\mathrm{D}}})\tilde{\psi}_{\mathrm{k}i'}\rangle - \epsilon_{\mathrm{k}i'} + \mu) \frac{1}{\sigma} f'\left(\frac{\epsilon_{\mathrm{k}i'}-\mu}{\sigma}\right) P^*_{\mathrm{k}i'i} P_{\mathrm{k}ji'} \\
&\quad - \frac{w_{\mathrm{k}} \sum_{i'=1}^{N} f'\left(\frac{\epsilon_{\mathrm{k}i'}-\mu}{\sigma}\right) P^*_{\mathrm{k}i'i} P_{\mathrm{k}ji'}}{\sum_{\mathrm{k}'\in\mathcal{K}} w_{\mathrm{k}'} \sum_{i'=1}^{N} f'\left(\frac{\epsilon_{\mathrm{k}'i'}-\mu}{\sigma}\right)} \sum_{\mathrm{k}'\in\mathcal{K}} w_{\mathrm{k}'} \sum_{i'=1}^{N} (\langle \tilde{\psi}_{\mathrm{k}'i'}, H_{\mathrm{k}'}(\rho_{\tilde{\Psi},\eta_{\mathrm{D}}})\tilde{\psi}_{\mathrm{k}'i'}\rangle - \epsilon_{\mathrm{k}'i'} + \mu) \frac{1}{\sigma} f'\left(\frac{\epsilon_{\mathrm{k}'i'}-\mu}{\sigma}\right) \\
&\quad + w_{\mathrm{k}} \sum_{i'\neq j'} \langle \tilde{\psi}_{\mathrm{k}j'}, H_{\mathrm{k}}(\rho_{\tilde{\Psi},\eta_{\mathrm{D}}})\tilde{\psi}_{\mathrm{k}i'}\rangle \frac{f_{\mathrm{k}j'} - f_{\mathrm{k}i'}}{\epsilon_{\mathrm{k}j'} - \epsilon_{\mathrm{k}i'}} P^*_{\mathrm{k}i'i} P_{\mathrm{k}jj'} \\
&= w_{\mathrm{k}} \Bigg(\sum_{i'=1}^{N} (\langle \tilde{\psi}_{\mathrm{k}i'}, H_{\mathrm{k}}(\rho_{\tilde{\Psi},\eta_{\mathrm{D}}})\tilde{\psi}_{\mathrm{k}i'}\rangle - \epsilon_{\mathrm{k}i'}) \frac{1}{\sigma} f'\left(\frac{\epsilon_{\mathrm{k}i'}-\mu}{\sigma}\right) P^*_{\mathrm{k}i'i} P_{\mathrm{k}ji'} \\
&\quad - \frac{\sum_{i'=1}^{N} f'\left(\frac{\epsilon_{\mathrm{k}i'}-\mu}{\sigma}\right) P^*_{\mathrm{k}i'i} P_{\mathrm{k}ji'}}{\sum_{\mathrm{k}'\in\mathcal{K}} w_{\mathrm{k}'} \sum_{i'=1}^{N} f'\left(\frac{\epsilon_{\mathrm{k}'i'}-\mu}{\sigma}\right)} \sum_{\mathrm{k}'\in\mathcal{K}} w_{\mathrm{k}'} \sum_{i'=1}^{N} (\langle \tilde{\psi}_{\mathrm{k}'i'}, H_{\mathrm{k}'}(\rho_{\tilde{\Psi},\eta_{\mathrm{D}}})\tilde{\psi}_{\mathrm{k}'i'}\rangle - \epsilon_{\mathrm{k}'i'}) \frac{1}{\sigma} f'\left(\frac{\epsilon_{\mathrm{k}'i'}-\mu}{\sigma}\right) \\
&\quad + \sum_{i'\neq j'} \langle \tilde{\psi}_{\mathrm{k}j'}, H_{\mathrm{k}}(\rho_{\tilde{\Psi},\eta_{\mathrm{D}}})\tilde{\psi}_{\mathrm{k}i'}\rangle \frac{f_{\mathrm{k}j'} - f_{\mathrm{k}i'}}{\epsilon_{\mathrm{k}j'} - \epsilon_{\mathrm{k}i'}} P^*_{\mathrm{k}i'i} P_{\mathrm{k}jj'} \Bigg).
\end{aligned}
$$

Here $\tilde{\Psi} = (\tilde{\Psi}_{\mathrm{k}})_{\mathrm{k}\in\mathcal{K}}$, $\eta_{\mathrm{D}} = (\eta_{\mathrm{k},\mathrm{D}})_{\mathrm{k}\in\mathcal{K}}$, $\tilde{\Psi}_{\mathrm{k}} = (\tilde{\psi}_{\mathrm{k}1}, \ldots, \tilde{\psi}_{\mathrm{k}N}) = \Psi_{\mathrm{k}} P_{\mathrm{k}}$, $\eta_{\mathrm{k},\mathrm{D}} \coloneqq \operatorname{Diag}(\epsilon_{\mathrm{k}1}, \ldots, \epsilon_{\mathrm{k}N})$, and

$$
\frac{f_{\mathrm{k}j'} - f_{\mathrm{k}i'}}{\epsilon_{\mathrm{k}j'} - \epsilon_{\mathrm{k}i'}} = \frac{1}{\sigma} f'\left(\frac{\epsilon_{\mathrm{k}i'}-\mu}{\sigma}\right)
$$

provided $\epsilon_{\mathrm{k}j'} = \epsilon_{\mathrm{k}i'}$.

When all η_{k} are diagonal matrix, we see from $P_{\mathrm{k}} = I_N$ for all $\mathrm{k} \in \mathcal{K}$ that

$$\begin{aligned}\frac{\partial \mathcal{F}}{\partial \eta_{\mathrm{k}ij}} = w_{\mathrm{k}}\Bigg(& (\langle \psi_{\mathrm{k}i}, H_{\mathrm{k}}(\rho_{\Psi,\eta})\psi_{\mathrm{k}i}\rangle - \epsilon_{\mathrm{k}i})\frac{1}{\sigma} f'\left(\frac{\epsilon_{\mathrm{k}i}-\mu}{\sigma}\right)\delta_{ij} \\ & - \frac{f'\left(\frac{\epsilon_{\mathrm{k}'i}-\mu}{\sigma}\right)\delta_{ij}}{\sum_{\mathrm{k}'} w_{\mathrm{k}'} \sum_{i'=1}^{N} f'\left(\frac{\epsilon_{\mathrm{k}'i'}-\mu}{\sigma}\right)} d_\mu \\ & + \langle \psi_{\mathrm{k}j}, H(\rho_{\Psi,\eta})\psi_{\mathrm{k}i}\rangle \frac{f_{\mathrm{k}j} - f_{\mathrm{k}i}}{\epsilon_{\mathrm{k}j} - \epsilon_{\mathrm{k}i}}(1-\delta_{ij})\Bigg)\end{aligned}$$

for any $\mathrm{k} \in \mathcal{K}$, where

$$d_\mu = \sum_{\mathrm{k}' \in \mathcal{K}} w_{\mathrm{k}'} \sum_{i'=1}^{N} (\langle \psi_{\mathrm{k}'i'}, H_{\mathrm{k}'}(\rho_{\Psi,\eta})\psi_{\mathrm{k}'i'}\rangle - \epsilon_{\mathrm{k}'i'})\frac{1}{\sigma} f'\left(\frac{\epsilon_{\mathrm{k}'i'}-\mu}{\sigma}\right). \tag{A.5}$$

We denote by $\nabla_{\eta_{\mathrm{k}}}\mathcal{F} = \mathcal{F}_{\eta_{\mathrm{k}}}^T = \left(\left(\frac{\partial \mathcal{F}}{\partial \eta_{\mathrm{k}ij}}\right)_{i,j=1}^{N}\right)^T$, $\nabla_\eta \mathcal{F} = (\nabla_{\eta_{\mathrm{k}}}\mathcal{F})_{\mathrm{k}\in\mathcal{K}}$.

Appendix B. Kohn-Sham equation. In this appendix, we show the associated standard Kohn-Sham equation for the ensemble Kohn-Sham DFT.

Let $\mathcal{L}_\Psi(\Phi, \eta, \Lambda) = 0$, i.e.,

$$H_{\mathrm{k}}(\rho_{\Phi,\eta})\Phi_{\mathrm{k}} F_{\eta_{\mathrm{k}}} = \mathcal{B}\Phi_{\mathrm{k}}\Lambda_{\mathrm{k}}, \quad \forall \mathrm{k} \in \mathcal{K}. \tag{B.1}$$

Thus we have

$$\Sigma_{\Phi_{\mathrm{k}},\eta_{\mathrm{k}}} F_{\eta_{\mathrm{k}}} = \Lambda_{\mathrm{k}}, \tag{B.2}$$

where $\Sigma_{\Phi_{\mathrm{k}},\eta_{\mathrm{k}}} = \langle \Phi_{\mathrm{k}}^* H(\rho_{\Phi,\eta})\Phi_{\mathrm{k}}\rangle$. Let $\mathcal{L}_\eta(\Phi, \eta, \Lambda) = 0$. Without loss of generality, let all η_{k} be diagonal. If not, by (3.7), we still have $\mathcal{L}_{\Psi_{\mathrm{k}}} = 0$ and $\mathcal{L}_{\eta_{\mathrm{k}}} = 0$ after diagonalizing η_{k} and then rotating the Φ_{k} and performing a similarity transformation on Λ_{k} accordingly.

Denote $\eta_{\mathrm{k}} = \mathrm{Diag}(\epsilon_{\mathrm{k}1}, \ldots, \epsilon_{\mathrm{k}N})$. Since f is strictly monotonic decreasing, the derivatives of f are always less than 0. We obtain from η_{k} being diagonal and $\mathcal{L}_{\eta_{\mathrm{k}}}(\Phi, \eta, \Lambda) = 0$ that $\Sigma_{\Phi_{\mathrm{k}},\eta_{\mathrm{k}}} = \eta_{\mathrm{k}} + cI$ is diagonal, where

$$c = \frac{d_\mu}{\frac{1}{\sigma}\sum_{\mathrm{k}} w_{\mathrm{k}} \sum_{i'=1}^{N} f'\left(\frac{\epsilon_{\mathrm{k}i'}-\mu}{\sigma}\right)}. \tag{B.3}$$

Here d_μ is defined as (A.5). Denote $\varepsilon_{\mathrm{k}i} = \epsilon_{\mathrm{k}i} + c$, then $\Sigma_{\Phi_{\mathrm{k}},\eta_{\mathrm{k}}} = \mathrm{Diag}(\varepsilon_{\mathrm{k}1}, \ldots, \varepsilon_{\mathrm{k}N})$. Consequently, we arrive at the standard Kohn-Sham equation

$$H(\rho)\phi_{\mathrm{k}i} = \varepsilon_{\mathrm{k}i}\mathcal{B}\phi_{\mathrm{k}i}, \quad i = 1, 2, \ldots, N. \tag{B.4}$$

where $\rho = \sum_{\mathrm{k}\in\mathcal{K}} w_{\mathrm{k}} \,\mathrm{tr}((\Phi_{\mathrm{k}}^*\Psi_{\mathrm{k}} + \langle \Phi_{\mathrm{k}}^* M\rangle \mathcal{Q}\langle M^*\Phi_{\mathrm{k}}\rangle)F_{\eta_{\mathrm{k}}})$, $\eta_{\mathrm{k}} = \mathrm{Diag}(\varepsilon_{\mathrm{k}1}, \varepsilon_{\mathrm{k}2}, \ldots, \varepsilon_{\mathrm{k}N})$.

If Λ_{k} are forced to be Hermitian, then we can derive the Kohn-Sham equation without the condition $\mathcal{L}_\eta(\Phi, \eta, \Lambda) = 0$. Indeed, it is clear that $\Sigma_{\Phi_{\mathrm{k}},\eta_{\mathrm{k}}} = \langle \Phi_{\mathrm{k}}^* H(\rho_{\Phi,\eta})\Phi_{\mathrm{k}}\rangle$ are Hermitian since Hamiltonian operator $H(\rho_{\Phi,\eta})$ is self-adjoint. It follows from $\Lambda_{\mathrm{k}}^* = \Lambda_{\mathrm{k}}$ and $F_{\eta_{\mathrm{k}}}^* = F_{\eta_{\mathrm{k}}}$ that

$$\Sigma_{\Phi_{\mathrm{k}},\eta_{\mathrm{k}}} F_{\eta_{\mathrm{k}}} = F_{\eta_{\mathrm{k}}} \Sigma_{\Phi_{\mathrm{k}},\eta_{\mathrm{k}}}. \tag{B.5}$$

Thus there exists $P \in \left(\mathcal{O}^{N\times N}\right)^{|\mathcal{K}|}$ such that

$$\Sigma_{\Phi_{\rm k}P_{\rm k},P_{\rm k}^*\eta_{\rm k}P_{\rm k}} = P_{\rm k}^*\Sigma_{\Phi_{\rm k},\eta_{\rm k}}P_{\rm k}, \quad F_{P_{\rm k}^*\eta P_{\rm k}} = P_{\rm k}^* F_{\eta_{\rm k}} P_{\rm k}, \quad P_{\rm k}^*\Lambda_{\rm k} P_{\rm k}$$

are diagonal. Let $\mathrm{Diag}(\varepsilon_{\rm k1},\ldots,\varepsilon_{{\rm k}N}) = P_{\rm k}^*\Lambda_{\rm k}F_{\eta_{\rm k}}^{-1}P_{\rm k}$. We still denote $\Phi_{\rm k}P_{\rm k}$ and $P_{\rm k}^*\eta_{\rm k}P_{\rm k}$ by $\Phi_{\rm k}$ and $\eta_{\rm k}$, respectively. Consequently, we arrive at (B.4).

The Kohn-Sham equations (B.4) are usually solved by the SCF iterations which is stated as Algorithm B.1.

Algorithm B.1 The SCF iteration method for solving ensemble Kohn-Sham DFT

1: Given $\epsilon > 0$, σ and initial guess of the input density $\rho_{\rm in}$. Set $\rho_{\rm out} = 0$;
2: **while** $\|\rho_{\rm out} - \rho_{\rm in}\| > \epsilon$ **do**
3: Obtain the input density $\rho_{\rm in}$ by some mixing schemes from $\rho_{\rm out}$ and the density of previous steps;
4: Solve the linear eigenvalue problems

$$H(\rho_{\rm in})\phi_{{\rm k}i} = \varepsilon_{{\rm k}i}\phi_{{\rm k}i},$$

to get eigenpairs $(\phi_{{\rm k}i},\varepsilon_{{\rm k}i})$, ${\rm k}\in\mathcal{K}$, $i = 1,2\ldots,N$;
5: Calculate μ and occupation numbers $f_{{\rm k}i}$ corresponding to eigenfunctions $\phi_{{\rm k}i}$ such that $\sum_{{\rm k}\in\mathcal{K}} w_{\rm k}\sum_{i=1}^{N} f_{{\rm k}i} = N_e$ and

$$f_{{\rm k}i} = f\left(\frac{\varepsilon_{{\rm k}i}-\mu}{\sigma}\right);$$

6: Calculate output density

$$\rho_{\rm out} = \sum_{{\rm k}\in\mathcal{K}} w_{\rm k}\,\mathrm{tr}((\Psi_{\rm k}^*\Psi_{\rm k} + \langle\Psi_{\rm k}^* M\rangle\mathcal{Q}\langle M^*\Psi_{\rm k}\rangle)F_{\eta_{\rm k}}),$$

where $F_{\eta_{\rm k}} = \mathrm{Diag}(f_{\rm k1}, f_{\rm k2},\ldots,f_{{\rm k}N})$;
7: **end while**

Acknowledgments. The authors would like to thank Professor Zhigang Wang for providing the configurations of the gold clusters, Professor Nicola Marzari for providing the configurations of the multicomponent systems, and Dr. Liwei Zhang for his helpful discussions.

REFERENCES

[1] K. Baarman, V. Havu, and T. Eirola, *Direct minimization for ensemble electronic structure calculations*, J. Sci. Comput., 66 (2016), pp. 1218–1233.
[2] A. D. Becke, *Perspective: Fifty years of density-functional theory in chemical physics*, J. Chem. Phys., 140 (2014), p. 18A301.
[3] P. E. Blöchl, *Projector augmented-wave method*, Phys. Rev. B, 50 (1994), pp. 17953–17979.
[4] C. L. Bris, ed., *Special Volume: Computational Chemistry*, vol. X of Handbook of Numerical Analysis, North-Holland, 2003.
[5] J. Callaway and N. March, *Density functional methods: Theory and applications*, in Solid State Physics, vol. 38, Elsevier, 1984, pp. 135–221.
[6] H. Chen, X. Gong, L. He, Z. Yang, and A. Zhou, *Numerical analysis of finite dimensional approximations of Kohn-Sham models*, Adv. Comput. Math., 38 (2013), pp. 225–256.

[7] H. Chen, X. Gong, and A. Zhou, *Numerical approximations of a nonlinear eigenvalue problem and applications to a density functional model*, Math. Methods Appl. Sci., 33 (2010), pp. 1723–1742.
[8] J. B. Conway, *A Course in Functional Analysis*, Springer, New York; London, 2007.
[9] X. Dai, Z. Liu, L. Zhang, and A. Zhou, *A conjugate gradient method for electronic structure calculations*, SIAM J. Sci. Comput., 39 (2017), pp. A2702–A2740.
[10] X. Dai, L. Zhang, and A. Zhou, *Adaptive step size strategy for orthogonality constrained line search methods*, arXiv: 1906.02883, (2020), pp. 1–24.
[11] Y. H. Dai and Y. Yuan, *A Nonlinear Conjugate Gradient Method with a Strong Global Convergence Property*, SIAM J. Optim., 10 (1999), pp. 177–182.
[12] C. Elsässer, M. Fähnle, C. T. Chan, and K. M. Ho, *Density-functional energies and forces with Gaussian-broadened fractional occupations*, Phys. Rev. B, 49 (1994), pp. 13975–13978.
[13] R. Fletcher and C. M. Reeves, *Function minimization by conjugate gradients*, Comput. J., 7 (1964), pp. 149–154.
[14] C. Freysoldt, S. Boeck, and J. Neugebauer, *Direct minimization technique for metals in density functional theory*, Phys. Rev. B, 79 (2009), p. 241103.
[15] C. L. Fu and K. M. Ho, *First-principles calculation of the equilibrium ground-state properties of transition metals: Applications to Nb and Mo*, Phys. Rev. B, 28 (1983), pp. 5480–5486.
[16] B. Gao, X. Liu, X. Chen, and Y.-x. Yuan, *A new first-order algorithmic framework for optimization problems with orthogonality constraints*, SIAM J. Optim., 28 (2018), pp. 302–332.
[17] M. J. Gillan, *Calculation of the vacancy formation energy in aluminium*, J. Phys.: Condens. Matter, 1 (1989), pp. 689–711.
[18] M. P. Grumbach, D. Hohl, R. M. Martin, and R. Car, *Ab initio molecular dynamics with a finite-temperature density functional*, J. Phys.: Condens. Matter, 6 (1994), p. 1999.
[19] M. F. Herbst and A. Levitt, *Black-box inhomogeneous preconditioning for self-consistent field iterations in density functional theory*, J. Phys.: Condens. Matter, 33 (2021), p. 085503.
[20] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bur. Standards, 49 (1952), pp. 409–436.
[21] S. Ismail-Beigi and T. Arias, *New algebraic formulation of density functional calculation*, Comput. Phys. Commun., 128 (2000), pp. 1–45.
[22] G. Kresse and J. Furthmüller, *Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set*, Comput. Mater. Sci., 6 (1996), pp. 15–50.
[23] K. Kreutz-Delgado, *The complex gradient operator and the CR-calculus*, arXiv: 0906.4835, (2009), pp. 1–74.
[24] L. Lin and C. Yang, *Elliptic preconditioner for accelerating the self-consistent field iteration in Kohn–Sham density functional theory*, SIAM J. Sci. Comput., 35 (2013), pp. S277–S298.
[25] R. M. Martin, *Electronic Structure: Basic Theory and Practical Methods*, Cambridge University Press, Cambridge, United Kingdom; New York, NY, second edition ed., 2020.
[26] N. Marzari, *Ab-Initio Molecular Dynamics for Metallic Systems*, PhD thesis, University of Cambridge, 1996.
[27] N. Marzari, D. Vanderbilt, A. De Vita, and M. C. Payne, *Thermal Contraction and Disordering of the Al(110) Surface*, Phys. Rev. Lett., 82 (1999), pp. 3296–3299.
[28] N. Marzari, D. Vanderbilt, and M. C. Payne, *Ensemble density-functional theory for* ab initio *molecular dynamics of metals and finite-temperature insulators*, Phys. Rev. Lett., 79 (1997), pp. 1337–1340.
[29] M. Methfessel and A. T. Paxton, *High-precision sampling for Brillouin-zone integration in metals*, Phys. Rev. B, 40 (1989), pp. 3616–3621.
[30] R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules*, no. 16 in International Series of Monographs on Chemistry, Oxford University Press, New York, 1994.
[31] E. Polak and G. Ribière, *Note sur la convergence de méthodes de directions conjuguées*, Rev. Francaise Informat Recherche Opertionelle, 16 (1969), pp. 35–43.
[32] B. Polyak, *The conjugate gradient method in extremal problems*, USSR Comp. Math. and Math. Phys., 9 (1969), pp. 94–112.
[33] *Quantum ESPRESSO*. https://www.quantum-espresso.org/.
[34] Á. Ruiz-Serrano and C.-K. Skylaris, *A variational method for density functional theory calculations on metallic systems with thousands of atoms*, J. Chem. Phys., 139 (2013), p. 054107.
[35] R. Schneider, T. Rohwedder, A. Neelov, and J. Blauert, *Direct minimization for calculating invariant subspaces in density functional computations of the electronic structure*,

J. Comput. Math., 27 (2009), pp. 360–387.
[36] N. Troullier and J. L. Martins, *Efficient pseudopotentials for plane-wave calculations*, Phys. Rev. B, 43 (1991), pp. 1993–2006.
[37] M. Ulbrich, Z. Wen, C. Yang, D. Klöckner, and Z. Lu, *A proximal gradient method for ensemble density functional theory*, SIAM J. Sci. Comput., 37 (2015), pp. A1975–A2002.
[38] D. Vanderbilt, *Soft self-consistent pseudopotentials in a generalized eigenvalue formalism*, Phys. Rev. B, 41 (1990), pp. 7892–7895.
[39] H. Zhang and W. W. Hager, *A nonmonotone line search technique and its application to unconstrained optimization*, SIAM J. Optim., 14 (2004), pp. 1043–1056.
[40] X. Zhang, J. Zhu, Z. Wen, and A. Zhou, *Gradient type optimization methods for electronic structure calculations*, SIAM J. Sci. Comput., 36 (2014), pp. 265–289.
[41] Z. Zhao, Z.-J. Bai, and X.-Q. Jin, *A Riemannian Newton algorithm for nonlinear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 752–774.
[42] Y. Zhou, H. Wang, Y. Liu, X. Gao, and H. Song, *Applicability of Kerker preconditioning scheme to the self-consistent density functional theory calculations of inhomogeneous systems*, Phys. Rev. E, 97 (2018), p. 033305.