Multiscale Model. Simul.
Vol. 21, No. 3, pp. 777–803

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 arising in 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 of the energy functional and the existence of the minimizer of the ensemble Kohn–Sham model. We propose an adaptive two-parameter step size strategy and the corresponding preconditioned conjugate gradient methods to solve the energy minimization model. Under some mild but reasonable assumptions, we prove the global convergence for the gradients of the energy functional produced by 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

MSC codes. 65K10, 65N25, 49S05, 35P30

DOI. 10.1137/22M1472103

1. Introduction.

The Kohn–Sham density functional theory (DFT) is widely used in electronic structure calculations [2, 22, 25, 31]. The underlying mathematical model is often formulated as either a nonlinear eigenvalue problem or an energy minimization problem with a 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 (see, e.g., [5, 6, 25, 31] and references therein). Recently, people have paid more attention to investigating the constrained energy minimization problem (see, e.g., [9, 38, 43, 44] 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

[*] Received by the editors January 18, 2022; accepted for publication (in revised form) January 30, 2023; published electronically July 12, 2023.
https://doi.org/10.1137/22M1472103

Funding: 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, the National Center for Mathematics and Interdisciplinary Sciences, CAS, and the CAS President's International Fellowship for Visiting Scientists under grant 2019VMA0029. The second author also acknowledges support from the European Union's Horizon 2020 research and innovation program (grant 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@lsec.cc.ac.cn, 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).

absent. More precisely, the classical Kohn–Sham DFT model becomes ill-posed due to the difficulty of separating 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 have been applied in computations. For instance, the ensemble Kohn–Sham DFT (or the finite-temperature Kohn–Sham DFT) has been developed (see, e.g., [23]), in which the associated total energy is a nonlinear functional of Kohn–Sham orbitals and pseudo-eigenvalues (or occupation numbers). We see that the ensemble Kohn–Sham DFT model 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 note that some preconditioners have also been constructed to accelerate the SCF iterations [19, 23, 24, 45]. Unfortunately, the convergence of the SCF iterations for the ensemble Kohn–Sham DFT model is not guaranteed in theory.

In the context of solving the constrained energy minimization problem arising in the ensemble Kohn–Sham DFT model, unlike the case of the classical Kohn–Sham DFT model, we need to treat the occupation numbers as additional variables. There are more challenges for designing and analyzing an efficient algorithm. For example, the energy of the classical Kohn–Sham DFT model is invariant under the unitary transformation of Kohn–Sham orbitals.[1] Thus, if Ψ is the ground state of the classical Kohn–Sham DFT, then $\{\Psi P : P \text{ is a unitary matrix}\}$ are the ground states, too. However, the unitary invariance of the energy functional for the ensemble Kohn–Sham DFT model is not clear. If Ψ is a ground state of the ensemble Kohn–Sham DFT model and P is a unitary matrix, ΨP may not be a ground state. As a result, it is necessary to take into account the unitary transformation and to apply an appropriate unitary transformation when designing an optimization algorithm for the ensemble Kohn–Sham DFT model. We refer the reader to [16, 17, 23] for constructing the unitary transformation for the Kohn–Sham orbitals. Ismail-Beigi and Arias [21] suggested expressing the unitary transformation as $P = e^{\mathrm{i}B}$ with B the Hermitian matrix and minimizing the energy functional with respect to B. However, the unitary transformation is incorporated into the model when some matrix representations relevant to the occupation numbers are applied. Marzari, Vanderbilt, and Payne [28] proposed an optimization algorithm by adopting a matrix representation of the occupation numbers, which we call the occupation matrix. They got a unitarily invariant functional of Kohn–Sham orbitals by minimizing the energy functional with respect to the occupation matrix, with which it is not necessary to construct the unitary transformation. Later on, Freysoldt, Boeck, and Neugebauer [14] introduced the so-called pseudo-Hamitonian matrix and proposed a preconditioned conjuagte gradient algorithm to minimize the energy functional with respect to the Kohn–Sham orbitals and the pseudo-Hamiltonian matrix simultaneously, in which the unitary transformation is constructed automatically by minimizing the energy functional with respect to the pseudo-Hamiltonian matrix. We mention that there is no theoretical numerical analysis for the above methods. Recently, Ulbrich et al. [40] studied a proximal gradient method for the ensemble Kohn–Sham DFT model but only for the Fermi–Dirac smearing. We also refer the reader to [1, 37] for more works on the direct minimization algorithms for the ensemble Kohn–Sham DFT model. To the best of our knowledge, there is little mathematical analysis on the ensemble Kohn–Sham DFT model and its

[1] We call it the unitary invariance of the energy functional of the classical Kohn–Sham DFT.

approximations. In this paper, we investigate the energy minimization model arising in the ensemble Kohn–Sham DFT from a mathematical viewpoint and design and analyze the associated optimization algorithms.

The rest of this paper is organized as follows. In next section, we introduce some basic notation and the energy minimization model arising in the ensemble Kohn–Sham DFT with the pseudo-eigenvalue matrix and the general smearing method. In section 3, we study the invariance of the energy functional and the existence of the minimizer of the ensemble Kohn–Sham DFT model. In section 4, we propose an adaptive two-parameter 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 for the gradients of the energy functional produced by our adaptive two-parameter step size based algorithms. 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. Due to space constraints, some detailed discussions, calculations, and proofs, which are technically needed but not essential to understand the main ideas, are provided in the self-contained supplement addition (supplement.pdf [local/web 9.53MB]) to this paper.

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 $\zeta_1, \zeta_2, \zeta_3 \in \mathbb{R}^3$ satisfy $\xi_i \cdot \zeta_j = 2\pi\delta_{ij}$, $i, j = 1, 2, 3$.

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

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

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

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

For $s \in \mathbb{R}$, we define the Sobolev space of complex-valued $\mathcal{R}$-periodic functions as

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

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

$$\Psi\Phi^* = \sum_{i=1}^N \psi_i \phi_i^*, \ \Psi^*\Phi = (\psi_i^*\phi_j)_{i,j=1}^N$$

and

$$A\Psi^* = \left(\sum_{j=1}^N a_{1j}\psi_j^*, \dots, \sum_{j=1}^N a_{Nj}\psi_j^*\right)^T, \ \Psi A = \left(\sum_{i=1}^N a_{i1}\psi_i, \dots, \sum_{i=1}^N a_{iN}\psi_i\right)$$

for any $A = (a_{ij})_{i,j=1}^N \in \mathbb{C}^{N\times N}$. We shall use the notation

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

We see that $\langle \Psi^*\Phi\rangle = \int_\Omega \Psi^*\Phi$. Denote the inner product as $\langle \Psi, \Phi\rangle = \mathrm{tr}\langle \Psi^*\Phi\rangle$, whose induced norm is $\|\Psi\| = \sqrt{\langle \Psi, \Psi\rangle}$.

For any $A, B \in \mathbb{C}^{N\times N}$, we define their inner product as $\langle A, B\rangle = \mathrm{tr}(A^*B)$, which induces the Frobenius norm $\|\cdot\|_F$ for $\mathbb{C}^{N\times N}$. We shall use the notation $\|\cdot\|_{sF}$ which is defined as $\|A\|_{sF} = \min_{c\in\mathbb{C}} \|cI_N - A\|_F$. It is easy to obtain $\|A\|_{sF} = \left\|\frac{\mathrm{tr}\, A}{N} I_N - A\right\|_F$.

The generalized Stiefel manifold associated with $\mathcal{B}$ is defined by

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

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

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

be the sets of unitary and Hermitian matrices in $\mathbb{C}^{N\times N}$, respectively.

We then introduce some projections, which will be applied in our algorithms. 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. We define a linear operator from $(H^1_\#(\Omega, \mathbb{C}))^N$ to K_Ψ by

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

which is a projection due to $P_\Psi^2 = P_\Psi$. Then the adjoint operator of P_Ψ is

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

We see that $P_\Psi^*(\Phi) \in \mathcal{T}_\Psi \mathcal{M}^N_{\mathcal{B},\mathbb{C}}$ and $P_\Psi(\Phi)$ 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 DFT, in which we adopt the matrix representation of occupation numbers [14, 28]. We see from Bloch's theorem [25] that the kinetic energy and the electronic density are given by integrals over the Brillouin zone (BZ).

If the BZ sampling is used to discretize the integrals, the ensemble Kohn–Sham energy functional with a general smearing approach can be formulated as

$$\mathcal{F}((\Psi_{\mathrm{k}})_{\mathrm{k}\in\mathcal{K}},(\eta_{\mathrm{k}})_{\mathrm{k}\in\mathcal{K}})=\mathcal{E}((\Psi_{\mathrm{k}})_{\mathrm{k}\in\mathcal{K}},(\eta_{\mathrm{k}})_{\mathrm{k}\in\mathcal{K}})-\sigma\sum_{k\in\mathcal{K}}w_{\mathrm{k}}\operatorname{tr}S\left(\frac{1}{\sigma}(\eta_{\mathrm{k}}-\mu I_N)\right) \tag{2.1}$$

with the Kohn–Sham orbitals $\Psi_{\mathrm{k}}\in(H^1_{\#}(\Omega,\mathbb{C}))^N$ and the matrix $\eta_{\mathrm{k}}\in\mathcal{S}^{N\times N}_{\mathbb{C}}$ (we call η_{k} the pseudo-eigenvalue matrix) for any $\mathrm{k}\in\mathcal{K}$, where S is a function associated with the entropy,

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

Here $\mathcal{K}$ is a finite subset of points in the BZ, $|\mathcal{K}|$ is the cardinality of $\mathcal{K}$, w_{k} is the weight relevant to the k-point $\mathrm{k}\in\mathcal{K}$ satisfying $\sum_{\mathrm{k}\in\mathcal{K}}w_{\mathrm{k}}=2$, N is the number of Kohn–Sham orbitals for each k-point, $\sigma=k_BT$ 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 smooth approximation to the step function which is sometimes called the smearing function, and the chemical potential μ is a function of η which will be determined later. The electronic density ρ is

$$\rho=\sum_{\mathrm{k}\in\mathcal{K}}w_{\mathrm{k}}\operatorname{tr}((\Psi^*_{\mathrm{k}}\Psi_{\mathrm{k}}+\langle\Psi^*_{\mathrm{k}}M\rangle\mathcal{Q}\langle M^*\Psi_{\mathrm{k}}\rangle)F_{\eta_{\mathrm{k}}}), \tag{2.2}$$

where $M=(\varphi_1,\dots,\varphi_K)\in(L^2_{\#}(\Omega,\mathbb{C}))^K$ and $\mathcal{Q}=(\mathcal{Q}_{ij})^K_{i,j=1}\in(L^2_{\#}(\Omega,\mathbb{C}))^{K\times K}$, satisfying $\mathcal{Q}^*=\mathcal{Q}$, are the pseudopotential projection states and the ultrasoft charge density contributions, respectively. Here K is the number of the pseudopotential projection states. $V_{\mathrm{loc}}\in L^2_{\#}(\Omega,\mathbb{C})$ is the local part of the pseudopotential, V_{nl} is the nonlocal part of the pseudopotential defined by $\Psi_{\mathrm{k}}\mapsto V_{\mathrm{nl}}(\Psi_{\mathrm{k}})=MD\langle M^*\Psi_{\mathrm{k}}\rangle$ with $D\in\mathcal{S}^{K\times K}_{\mathbb{C}}$, and $\mathcal{E}_{\mathrm{xc}}$ is the exchange-correction functional. Note that the form of (2.1) is suitable for the full potential calculations, the pseudopotential approximations [39, 41], and the projector augmented wave (PAW) method [3]. For instance, if the norm-conserving pseudopotential is applied, then $\mathcal{Q}=0$ and $\rho=\sum_{\mathrm{k}\in\mathcal{K}}w_{\mathrm{k}}\operatorname{tr}(\Psi^*_{\mathrm{k}}\Psi_{\mathrm{k}}F_{\eta_{\mathrm{k}}})$. 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 $N>N_b:=N_e/2$, where N_e is the number of electrons. For example, in Quantum ESPRESSO [35], N is set as $\max(\lfloor 1.2N_b\rfloor,N_b+4)$ by default, where $\lfloor x\rfloor$ is the greatest integer not larger than x.

For simplicity, we here and hereafter assume that there is only one k-point in our following analysis; that is, we assume $\mathcal{K}=\{\mathrm{k}_0\}$. And we will use Ψ and η as the associated Kohn–Sham orbitals and the pseudo-eigenvalue matrix, respectively. Most of the analysis and description of the algorithms can be easily extended to the case of the multiple k-points, for which a self-contained description including some special statements to be taken into account is provided in the supplement (supplement.pdf [local/web 9.53MB]).

Throughout this paper, we assume that f and S satisfy some of the following properties:

A.I f and S are analytic functions on $\mathbb{R}$ and satisfy $S'(x)=xf'(x)$.
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 the assumptions above, we then address μ as a function of η. For any given $\eta\in\mathcal{S}_{\mathbb{C}}^{N\times N}$, we obtain from the above assumptions that there is one and only one $\mu\in\mathbb{R}$ satisfying $2\,\mathrm{tr}\,F_\eta=N_e$. More specifically, the assumptions A.I–A.II imply the existence of $\mu\in\mathbb{R}$ such that $2\,\mathrm{tr}\,F_\eta=N_e$ for any given $\eta\in\mathcal{S}_{\mathbb{C}}^{N\times N}$, which means that there exists a function $\mu(\eta)$ such that $2\,\mathrm{tr}\,F_\eta=N_e$. Further, if the assumption A.IV is satisfied, then μ is unique. Thus, μ in (2.1) is defined as the unique function of η from $\mathcal{S}_{\mathbb{C}}^{N\times N}$ to $\mathbb{R}$ such that $2\,\mathrm{tr}\,F_\eta=N_e$.

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

- The Fermi–Dirac smearing [4]:
$$f_{\mathrm{FD}}(x)=\frac{1}{1+e^x},\ S_{\mathrm{FD}}(x)=-[f_{\mathrm{FD}}(x)\ln f_{\mathrm{FD}}(x)+(1-f_{\mathrm{FD}}(x))\ln(1-f_{\mathrm{FD}}(x))].$$
- The Gaussian smearing [12, 15]: $f_{\mathrm{GS}}(x)=\frac{1}{2}(1-\mathrm{erf}(x))$, $S_{\mathrm{GS}}(x)=\frac{1}{2\sqrt{\pi}}e^{-x^2}$.
- The Methfessel–Paxton smearing [29]:
$$f_{\mathrm{MP},m}(x)=f_{\mathrm{GS}}(x)+\sum_{i=1}^{m}A_iH_{2i-1}(x)e^{-x^2},\ S_{\mathrm{MP},m}(x)=\frac{1}{2}A_mH_{2m}(x)e^{-x^2},$$
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].

Note that all the above smearing functions satisfy the assumptions A.I–A.III. We see that the Fermi–Dirac smearing and the Gaussian smearing satisfy the assumption A.IV. Thus, μ can be defined as the unique function of η when these two smearing approaches are applied. However, μ may not be unique for some other smearing functions, such as the Methfessel–Paxton smearing function and the Marzari–Vanderbilt smearing function, that do not satisfy the assumption A.IV. In practice, we will always assume that μ is a function of η such that $2\,\mathrm{tr}\,F_\eta=N_e$.

According to the ensemble Kohn–Sham DFT, we can obtain the ground state of the system by solving the following constrained minimization problem:

$$\inf_{(\Psi,\eta)\in\mathcal{M}^N_{\mathcal{B},\mathbb{C}}\times\mathcal{S}_{\mathbb{C}}^{N\times N}}\mathcal{F}(\Psi,\eta),\tag{2.3}$$

where $\mathcal{B}$ is an operator defined by $\mathcal{B}\Phi=\Phi+MQ\langle M^*\Phi\rangle$ with $Q=\int_\Omega\mathcal{Q}(r)\mathrm{d}r$ for any $\Phi\in(L^2_\#(\Omega,\mathbb{C}))^N$. Note that $\mathcal{B}$ is bounded and self-adjoint. The associated Lagrange functional is

$$\mathcal{L}(\Psi,\eta,\Lambda)=\mathcal{F}(\Psi,\eta)-2\,\mathrm{tr}[\Lambda^*(\langle\Psi^*\mathcal{B}\Psi\rangle-I_N)]\tag{2.4}$$

with the Lagrange multiplier $\Lambda \in \mathbb{C}^{N\times N}$. Note that the restriction $\eta \in \mathcal{S}_{\mathbb{C}}^{N\times N}$ has no contribution to (2.4) since our discussion with respect to η throughout this paper is in the linear space $\mathcal{S}_{\mathbb{C}}^{N\times N}$.

Assume that the functional $\mathcal{E}_{\mathrm{xc}}$ is differentiable. Since Ψ is complex-valued and $\mathcal{F}$ is real-valued, $\mathcal{F}$ is not differentiable with respect to Ψ. Here, we consider the Wirtinger derivatives [36], for which more discussions are provided in section SM1.3 of the supplement. Generally speaking, we regard Ψ and $\bar{\Psi} := (\Psi^*)^T$ as two independent variables and view $\mathcal{F}$ as a functional of Ψ, $\bar{\Psi}$, and η. Then we get

$$\mathcal{F}_\Psi(\Psi,\eta) = 2H(\rho)\Psi F_\eta \quad \text{and} \quad \mathcal{L}_\Psi(\Psi,\eta,\Lambda) = 2(H(\rho)\Psi F_\eta - \mathcal{B}\Psi\Lambda).$$

Here $\mathcal{F}_\Psi$ and $\mathcal{L}_\Psi$ are Wirtinger derivatives, $H(\rho) = -\frac{1}{2}(\mathrm{i}k_0 + \nabla)^2 + \tilde{V}_{\mathrm{loc}}(\rho) + \tilde{V}_{\mathrm{nl}}(\rho)$, $\tilde{V}_{\mathrm{loc}}(\rho) = V_{\mathrm{loc}} + \int_\Omega \frac{\rho(r)}{|\cdot - r|}\mathrm{d}r + V_{\mathrm{xc}}(\rho)$, $\tilde{V}_{\mathrm{nl}}(\rho) : \Psi \mapsto V_{\mathrm{nl}}(\Psi) + M\tilde{D}\langle M^*\Psi\rangle$, $V_{\mathrm{xc}}(\rho) = \frac{\delta\mathcal{E}_{\mathrm{xc}}}{\delta\rho}$, and $\tilde{D} = \int_\Omega \tilde{V}_{\mathrm{loc}}(\rho)(r)\mathcal{Q}(r)\,\mathrm{d}r \in \mathcal{S}_{\mathbb{C}}^{K\times K}$. We use the convenient notation $\mathcal{F}(\Psi,\eta) = \mathcal{F}(\Psi,\bar{\Psi},\eta)$ and $\mathcal{L}(\Psi,\eta,\Lambda) = \mathcal{L}(\Psi,\bar{\Psi},\eta,\Lambda)$.

Throughout this paper, we view $\nabla_\Psi\mathcal{F}(\Psi,\eta)$ and $\nabla_\eta\mathcal{F}(\Psi,\eta)$ with

$$\begin{aligned} &\nabla_\Psi\mathcal{F}(\Psi,\eta) := 2\mathcal{L}_\Psi(\Psi,\eta,\langle\Psi^* H(\rho)\Psi\rangle F_\eta) = 4(H(\rho)\Psi - \mathcal{B}\Psi\langle\Psi^* H(\rho)\Psi\rangle)F_\eta, \\ &\nabla_\eta\mathcal{F}(\Psi,\eta) := \mathcal{L}_\eta^T = \mathcal{F}_\eta^T \end{aligned} \tag{2.5}$$

as the gradients of $\mathcal{F}$ under the constraint $(\Psi,\eta) \in \mathcal{M}_{\mathcal{B},\mathbb{C}}^N \times \mathcal{S}_{\mathbb{C}}^{N\times N}$, where $\mathcal{F}_\eta = (\frac{\partial\mathcal{F}}{\partial\eta_{ij}})_{i,j=1}^N$. For convenience, we simply call them the gradients and still use the notation ∇_Ψ and ∇_η. Due to the space limitation, we here only provide the expression of $\frac{\partial\mathcal{F}}{\partial\eta_{ij}}$ when η is a diagonal matrix as

$$\frac{\partial\mathcal{F}}{\partial\eta_{ij}} = 2\left(\chi_{ii}(\langle\psi_i, H(\rho)\psi_i\rangle - \epsilon_i)\delta_{ij} - \frac{\chi_{ii}\delta_{ij}}{2\sum_{i'=1}^N \chi_{i'i'}}d_\mu + \chi_{ji}\langle\psi_j, H(\rho)\psi_i\rangle(1-\delta_{ij})\right),$$

where $\Psi = (\psi_1, \psi_2, \ldots, \psi_N)$, $\eta = \mathrm{Diag}(\epsilon_1, \epsilon_2, \ldots, \epsilon_N)$, $f_i = f((\epsilon_i - \mu)/\sigma)$,

$$\chi_{ij} = \begin{cases} \frac{f_i - f_j}{\epsilon_i - \epsilon_j} & \text{if } \epsilon_i \neq \epsilon_j, \\ \frac{1}{\sigma}f'\left(\frac{\epsilon_i - \mu}{\sigma}\right) & \text{if } \epsilon_i = \epsilon_j, \end{cases} \tag{2.6}$$

and

$$d_\mu = 2\sum_{i'=1}^N \chi_{i'i'}(\langle\psi_{i'}, H(\rho)\psi_{i'}\rangle - \epsilon_{i'}). \tag{2.7}$$

The detailed calculations for $\nabla_\eta\mathcal{F}$ with respect to the general $\eta \in \mathcal{S}^{N\times N}$ are provided in section SM1.3 of the supplement.

Let $\nabla\mathcal{F} = (\nabla_\Psi\mathcal{F}, \nabla_\eta\mathcal{F})$ and $\|\nabla\mathcal{F}\| = (\|\nabla_\Psi\mathcal{F}\|^2 + \|\nabla_\eta\mathcal{F}\|_{sF}^2)^{1/2}$. It is clear that $\mathcal{L}_\Psi(\Psi,\eta,\Lambda) = 0$ and $\mathcal{L}_\eta(\Psi,\eta,\Lambda) = 0$ mean that $\nabla\mathcal{F}(\Psi,\eta) = 0$. Conversely, $\nabla\mathcal{F}(\Psi,\eta) = 0$ means that there exists some Λ such that $\mathcal{L}_\Psi(\Psi,\eta,\Lambda) = 0$ and $\mathcal{L}_\eta(\Psi,\eta,\Lambda) = 0$. Thus, $\nabla\mathcal{F}(\Psi,\eta) = 0$ implies that (Ψ,η) is a stationary point of the problem (2.3). We see that any minimizer (Φ,η) of (2.3) satisfies $\mathcal{L}_\Psi(\Phi,\eta,\Lambda) = 0$ and $\mathcal{L}_\eta(\Phi,\eta,\Lambda) = 0$ for some $\Lambda \in \mathbb{C}^{N\times N}$, which lead to the following Kohn–Sham equation:

$$\begin{cases} H(\rho)\phi_i = \varepsilon_i\mathcal{B}\phi_i, & i = 1,2,\ldots,N, \\ \displaystyle\int_\Omega \phi_i^*\mathcal{B}\phi_j = \delta_{ij}, & i,j = 1,2,\ldots,N. \end{cases}$$

The details are shown in section SM1.4 of the supplement.

3. Mathematical analysis.

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

3.1. Invariance.

We first have the following invariance of the energy functional.

THEOREM 3.1. *If the assumptions* A.I, A.II, *and* A.IV *are satisfied, then it holds that*

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

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

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 $c \in \mathbb{R}$, $(\Psi, \eta) \in (H^1_\#(\Omega, \mathbb{C}))^N \times \mathcal{S}^{N\times N}_{\mathbb{C}}$, and $P \in \mathcal{O}^{N\times N}_{\mathbb{C}}$.

We first prove (3.2). For any given $\eta \in \mathcal{S}^{N\times N}_{\mathbb{C}}$, we see that there is a unique $\mu(\eta)$ that satisfies $2\,\mathrm{tr}\, F_\eta = N_e$. Thus, it follows from

$$2\,\mathrm{tr}\, f\left(\frac{\eta - \mu(\eta)I_N}{\sigma}\right) = N_e = 2\,\mathrm{tr}\, f\left(\frac{\eta + cI_N - \mu(\eta + cI_N)I_N}{\sigma}\right)$$

that $\mu(\eta + cI_N) = \mu(\eta) + c$ for any $c \in \mathbb{R}$. Therefore, we have $F_{\eta + cI_N} = F_\eta$ and

$$S\left(\frac{1}{\sigma}(\eta + cI_N - \mu(\eta + cI_N)I_N)\right) = S\left(\frac{1}{\sigma}(\eta - \mu(\eta)I_N)\right),$$

which lead to $\rho_{\Psi,\eta+cI_N} = \rho_{\Psi,\eta}$ and arrive at (3.2). Here we express ρ defined by (2.2) as $\rho_{\Psi,\eta}$ to express the relationship between the density and Ψ, η.

Next we prove the equation (3.3). Since f and S are analytic, we have

$$Pf(A)P^* = f(PAP^*),\ PS(A)P^* = S(PAP^*) \quad \forall A \in \mathcal{S}^{N\times N}_{\mathbb{C}},\ P \in \mathcal{O}^{N\times N}_{\mathbb{C}}.$$

Similarly, by the uniqueness of μ such that $2\,\mathrm{tr}\, F_\eta = N_e$, we get $\mu(P^*\eta P) = \mu(\eta)$ for any $P \in (\mathcal{O}^{N\times N}_{\mathbb{C}})^{\mathcal{K}}$. Note that

$$\begin{aligned}\rho_{\Psi P,\eta} &= 2\,\mathrm{tr}(P^*(\Psi^*\Psi + \langle \Psi^* M\rangle \mathcal{Q} \langle M^*\Psi\rangle)PF_\eta)\\ &= 2\,\mathrm{tr}((\Psi^*\Psi + \langle \Psi^* M\rangle \mathcal{Q} \langle M^*\Psi\rangle)F_{P\eta P^*})\\ &= \rho_{\Psi, P\eta P^*}.\end{aligned}$$

We have

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

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,

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

Finally, we obtain 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 see from the proof above that there always exists a function $\mu(\eta)$ such that $2\,\mathrm{tr}\, F_\eta = N_e$ and $\mu(\eta + cI_N) = \mu(\eta) + c$ for any $c \in \mathbb{R}$ under the assumptions A.I and A.II. We get from (3.1) that

$$\inf_{(\Psi,\eta)\in\mathcal{M}^N_{\mathcal{B},\mathbb{C}}\times\mathcal{D}^{N\times N}} \mathcal{F}(\Psi,\eta) = \inf_{(\Psi,\eta)\in\mathcal{M}^N_{\mathcal{B},\mathbb{C}}\times\mathcal{S}^{N\times N}_{\mathbb{C}}} \mathcal{F}(\Psi,\eta), \tag{3.4}$$

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

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

is the original ensemble Kohn–Sham DFT model, which means that the model (2.3) is equivalent to the original ensemble Kohn–Sham DFT model.

By a direct calculation, we have the following theorem for the gradients of $\mathcal{F}$.

THEOREM 3.2. *Suppose the assumptions* A.I*,* A.II*, and* A.IV *are satisfied. Given are* $c \in \mathbb{R}$*,* $(\Psi,\eta) \in (H^1_\#(\Omega,\mathbb{C}))^N \times \mathcal{S}^{N\times N}_{\mathbb{C}}$*, and* $P \in \mathcal{O}^{N\times N}_{\mathbb{C}}$*.*

1. *There hold*

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

2. $\nabla_\eta\mathcal{F}(\Psi,\eta)$ *is a Hermitian matrix;*
3. $\mathrm{tr}\,\nabla_\eta\mathcal{F}(\Psi,(\eta + cI_N)) = 0.$

We mention that the update step of our new algorithms in section 4 is based on the Theorems 3.1 and 3.2. In addition, in the supplement (supplement.pdf [local/web 9.53MB]), we also provide the extended versions of Theorems 3.1 and 3.2 in the case of the multiple k-points, together with further discussions.

3.2. Existence of the minimizer. In this subsection, we study the existence of the minimizer of the ensemble Kohn–Sham DFT model. We consider the case that the sampling of k-points is at Γ point ($\mathcal{K} = \{(0,0,0)\}$) only, for which Ψ, η, and other corresponding functions are real-valued. For the general sampling $\mathcal{K}$, the existence of the minimizer of the ensemble Kohn–Sham DFT model is still open.

Since only real values are taken into account in this subsection, we shall use the notation associated with real values by removing $\mathbb{C}$ and replacing the conjugate transpose symbol $*$ by the transpose symbol T in the notation introduced before.

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

$$\mathcal{N}\in\mathscr{P}(3,(c_1,c_2))\,(c_1\geq 0)\text{ or }\mathcal{N}\in\mathscr{P}(4/3,(c_1,c_2), \tag{3.5}$$

where

$$\mathscr{P}\,(p,(c_1,c_2))=\{g:\exists a_1,a_2\in\mathbb{R}\text{ such that }c_1t^p+a_1\leq g(t)\leq c_2t^p+a_2\quad\forall t\geq 0\}$$

with $c_1\in\mathbb{R}$ and $p,c_2\in[0,\infty)$. We assume that there exists a constant $\alpha>0$ such that the following inequality holds:

$$(\psi,\mathcal{B}\psi)\geq\alpha\|\psi\|^2\quad\forall\psi\in L^2_\#(\Omega). \tag{3.6}$$

As mentioned in section 2.2, we assume that the assumptions A.I–A.IV are satisfied. Let

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

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

We first have the following lemma.

Lemma 3.3. *If the assumptions* A.I, A.II, *and* A.IV *hold, then it holds that*

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

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

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

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

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 (3.4) yields the conclusion. □

Let $f(-\infty)=1,f(+\infty)=0$ and $S(-\infty)=\lim_{x\to-\infty}S(x),S(+\infty)=\lim_{x\to+\infty}S(x)$; then f and S are continuous on $[-\infty,+\infty]$ and $f([-\infty,\infty])=[0,1]$. We see from the assumption A.IV that the inverse function of f exists. Thus $S\circ f^{-1}$ is continuous on $[0,1]$. By Lemma 3.3, instead of $\inf_{(\Psi,F)\in\mathcal{M}_\mathcal{B}^N\times\mathcal{S}^{N\times N}}\mathcal{F}(\Psi,\eta)$, we consider the following minimization problem:

$$\inf_{(\Psi,F)\in\mathcal{M}_\mathcal{B}^N\times\overline{\mathscr{F}}_{\rm occ}}\widetilde{\mathcal{F}}(\Psi,F). \tag{3.7}$$

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

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

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

$$\widetilde{\widetilde{\mathcal{E}}}(\Psi F^{1/2})=\widetilde{\mathcal{E}}(\Psi,F),\forall(\Psi,F)\in\mathcal{M}_\mathcal{B}^N\times\overline{\mathscr{F}}_{\rm occ}. \tag{3.8}$$

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

PROPOSITION 3.4. *Suppose* (3.5) *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 (3.6), Jensen's inequality, and the similar arguments in [7], we get that $\widetilde{\mathcal{E}}(\Psi, F)$ is bounded below over $\mathcal{M}^N_{\mathcal{B}} \times \overline{\mathscr{F}}_{\text{occ}}$.

PROPOSITION 3.5. *If* (3.5) *and* (3.6) *hold, then there exist constants* $C > 0$ *and* $b > 0$ *such that*

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

Finally, we obtain the existence of a minimizer of (3.7).

THEOREM 3.6. *If* (3.5), (3.6)*, and the assumptions* A.I–A.IV *hold, then there exists* $(\Phi_*, F_*) \in \mathcal{M}^N_{\mathcal{B}} \times \overline{\mathscr{F}}_{\text{occ}}$ *such that*

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

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

Let $\{\Psi^{(n)}\}_{n\in\mathbb{N}} \subset \mathcal{M}^N_{\mathcal{B}}$ and $\{F^{(n)}\}_{n\in\mathbb{N}} \subset \overline{\mathscr{F}}_{occ}$ 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 on 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

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

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

$$F^{(n)} = \langle (\Psi^{(n)}(F^{(n)})^{1/2})^T \Psi^{(n)}(F^{(n)})^{1/2} \rangle \to F_*, \tag{3.9}$$

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 (3.8), (3.9), and Proposition 3.4, we obtain

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

This completes the proof. □

4. Numerical approximations.

Let

$$V_{N_G} = \operatorname{span}\left\{ e_{\mathrm{G}} : \mathrm{G} \in \mathcal{R}^*, \frac{1}{2}\left|\mathrm{k}_0 + \mathrm{G}\right|^2 \le E_{\text{cut}} \right\},$$

where E_{cut} is a given cutoff energy and N_G is the number of planewaves satisfying $\frac{1}{2}\left|\mathrm{k}_0 + \mathrm{G}\right|^2 \le E_{\text{cut}}$. A finite planewave discretization of the ensemble Kohn–Sham DFT minimization problem (2.3) reads as follows:

$$\inf_{(\Psi,\eta)\in\mathcal{M}^N_{\mathcal{B},\mathbb{C},N_G}\times\mathcal{S}^{N\times N}_{\mathbb{C}}} \mathcal{F}(\Psi,\eta), \tag{4.1}$$

where $\mathcal{M}^N_{\mathcal{B},\mathbb{C},N_G} = \{\Psi \in (V_{N_G})^N : \langle\Psi^*\mathcal{B}\Psi\rangle = I_N\}$. For the continuous system (2.3), we can only prove the existence of its minimizer for the case of the Γ point. However, for the discrete system (4.1) with any $\mathrm{k}_0 \in \mathrm{BZ}$, we are able to prove the existence of its minimizer by using the similar proof for Theorem 3.6 since $\mathcal{M}^N_{\mathcal{B},\mathbb{C},N_G}$ is compact. That is, the conclusion of Theorem 3.6 holds true for the discrete problem (4.1) with any $\mathrm{k}_0 \in \mathcal{K}$. In addition, the invariance of the energy functional addressed in section 3.1 also holds since $\mathcal{M}^N_{\mathcal{B},\mathbb{C},N_G} \subset (H^1_{\#}(\Omega,\mathbb{C}))^N$.

When we apply some line search based optimization methods, such as the gradient type method, the conjugate type method, and so on, to solve the minimization problem (4.1), the iterative behavior for Ψ and η may be different. So it is better to apply different step sizes for Ψ and η when we apply a line search based optimization method to solve (4.1). Inspired by the adaptive step size strategy proposed in [10], we propose an adaptive two-parameter step size strategy.

4.1. Adaptive two-parameter step size strategy.

An adaptive step size strategy can be concluded as the following four steps [10]:

Initialize → Estimate → Judge → Improve.

At the nth iteration, we first give some initial guess for the step sizes $(t_\Psi^{n,\text{initial}}, t_\eta^{n,\text{initial}})$. We require $t_\Psi^{n,\text{initial}} \ge t_\Psi^{\min}, t_\eta^{n,\text{initial}} \ge t_\eta^{\min}$ with $t_\Psi^{\min}, t_\eta^{\min}$ being some small positive constants to ensure that the initial step sizes are not too small. Then we introduce the other three steps for our adaptive two-parameter step size strategy one by one.

Let $D_\Psi^{(n)} \in \mathcal{T}_{\Psi^{(n)}}\mathcal{M}^N_{\mathcal{B},\mathbb{C},N_G}$, $D_\eta^{(n)} \in \mathcal{S}^{N\times N}_{\mathbb{C}}$. We use $\operatorname{ortho}(\Psi^{(n)}, D_\Psi^{(n)}, t_\Psi)$ to denote one step from $\Psi^{(n)} \in \mathcal{M}^N_{\mathcal{B},\mathbb{C},N_G}$ with the search direction $D_\Psi^{(n)}$ and the step size t_Ψ to the next point on $\mathcal{M}^N_{\mathcal{B},\mathbb{C},N_G}$. In our numerical experiments, we apply the QR strategy, which is defined by

$$\operatorname{ortho}_{\text{QR}}(\Psi, D_\Psi, t_\Psi) = (\Psi + t_\Psi D_\Psi)(L^*)^{-1}, \tag{4.2}$$

where L is the lower triangular matrix such that $LL^* = I_N + t_\Psi^2\langle D_\Psi^*\mathcal{B}D_\Psi\rangle$. We refer the reader to [9] for some other unitarity preserving strategies. For convenience, we simply denote $\mathcal{F}(\operatorname{ortho}(\Psi^{(n)}, D_\Psi^{(n)}, t_\Psi), \eta^{(n)} + t_\eta D_\eta^{(n)})$ by $\bar{\mathcal{F}}_n(t_\Psi, t_\eta)$. By a simple calculation, we have

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

We assume $\langle(D_\Psi^{(n)})^*\mathcal{B}\Psi^{(n)}\rangle = 0$ to ensure

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

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

$$\frac{\partial \bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0) \le 0, \ \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\eta}(0,0) \le 0, \quad n = 0,1,2,\dots, \tag{4.3}$$

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 $\nabla \mathcal{F}(\Psi^{(n)}, \eta^{(n)}) \neq 0$ in the adaptive two-parameter step size strategy; otherwise we have obtained a stationary point of the problem (4.1).

Estimate. The final step sizes are expected to satisfy the following condition:

$$\bar{\mathcal{F}}_n(t_\Psi^{(n)}, t_\eta^{(n)}) - \mathcal{C}_n \le \nu \left(t_\Psi^{(n)} \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0) + t_\eta^{(n)} \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\eta}(0,0) \right), \ n = 0,1,2,\dots, \tag{4.4}$$

where $\nu \in (0,1)$ is a given parameter. Here $\mathcal{C}_n$ can be chosen as that introduced in [42] as follows:

$$\begin{cases} \mathcal{C}_0 = \mathcal{F}(\Psi^{(0)}, \eta^{(0)}), \ Q_0 = 1, \\ Q_n = \alpha Q_{n-1} + 1, \\ \mathcal{C}_n = (\alpha Q_{n-1} \mathcal{C}_{n-1} + \mathcal{F}(\Psi^{(n)}, \eta^{(n)}))/Q_n, \end{cases} \tag{4.5}$$

where $\alpha \in [0,1)$ is a given parameter. We see that it is very expensive to calculate $\bar{\mathcal{F}}_n(t_\Psi^{(n)}, t_\eta^{(n)})$. Therefore, we consider the approximation of the energy functional $\mathcal{F}$ around $(\Psi^{(n)}, \eta^{(n)})$ as follows:

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

where $c_{n,1}, c_{n,2} \ge 0$ 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 (4.4) by the right-hand term of (4.6), we obtain

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

or, equivalently,

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

Hence, we propose the following estimator:

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

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

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

and then calculate the estimator $\zeta_n(t_\Psi^{(n)}, t_\eta^{(n)})$. We point out here that, for the case of multiple k-points, $\|D_\Psi^{(n)}\|$ and $\|D_\eta^{(n)}\|_{sF}$ will be replaced by $\|D_\Psi^{(n)}\|_\infty$ and $\|D_\eta^{(n)}\|_{sF,\infty}$, respectively, and all the derivation and analysis still hold true. The details are provided in section SM3.1 of the supplement.

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

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

then we accept this step sizes. Otherwise, we will improve $(t_\Psi^{(n)}, t_\eta^{(n)})$ by some strategy.

Improve. If $(t_\Psi^{(n)}, t_\eta^{(n)})$ is not accepted, then we set the minimizer of the approximation of $\bar{\mathcal{F}}_n$ (the right-hand side of (4.6)) to be the step size. Since the local approximation is applied in the neighborhood of $(\Psi^{(n)}, \eta^{(n)})$, we take

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

Here and hereafter, $-\frac{1}{c_{n,1}}\frac{\partial \bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0)$ will be 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 $-\frac{1}{c_{n,2}}\frac{\partial \bar{\mathcal{F}}_n}{\partial t_\eta}(0,0)$ will be 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 step sizes (4.8) always satisfy (4.7) if $\nu \in (0, 1/2]$. To ensure the convergence of the iterations in our following analysis, we do some adjustments to the above step sizes such that

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

where $\bar{c} > 1 > \underline{c} > 0$ are given constants. Specifically, if (4.9) is not satisfied, we then reduce one of the two step sizes to make it satisfy (4.9).

We summarize the above process as Algorithm 4.1, in which steps 6–10 are used to make sure that (4.9) is satisfied.

Note that it is very difficult to calculate the second derivatives of $\bar{\mathcal{F}}_n(t_\Psi, t_\eta)$. Here we provide a strategy to make approximations $c_{n,1}$ and $c_{n,2}$ as good as we can. Our idea is to get $c_{n,1}$ and $c_{n,2}$ by one trial step with step sizes $(t_\Psi^{\text{trial}}, t_\eta^{\text{trial}})$. We first set one trial step as

$$\begin{aligned} t_\Psi^{\text{trial}} &= \min\left(\max(t_\Psi^{\text{trial,min}}, t_\Psi^{(n-1)}), \frac{\theta_\Psi^{(n)}}{\|D_\Psi^{(n)}\|}\right),\\ t_\eta^{\text{trial}} &= \min\left(\max(t_\eta^{\text{trial,min}}, t_\eta^{(n-1)}), \frac{\theta_\eta^{(n)}}{\|D_\eta^{(n)}\|_{sF}}\right),\end{aligned} \tag{4.10}$$

Algorithm 4.1 Adaptive two-parameter step size strategy (ATPStep).

```
Input: $\Psi$, $\eta$, $D_\Psi$, $D_\eta$, $t_\Psi^{\text{initial}}$, $t_\eta^{\text{initial}}$, $t_\Psi^{\min}$, $t_\eta^{\min}$, $\nu$, $\underline{c}$, $\bar{c}$, $c_1$, $c_2$, $\theta_\Psi$, $\theta_\eta$, $\mathcal{C}$
Output: $t_\Psi$, $t_\eta$
Set
    $t_\Psi = \min\left(\max(t_\Psi^{\text{initial}}, t_\Psi^{\min}), \frac{\theta_\Psi}{\|D_\Psi\|}\right), \; t_\eta = \min\left(\max(t_\eta^{\text{initial}}, t_\eta^{\min}), \frac{\theta_\eta}{\|D_\eta\|_{sF}}\right);$
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 $\bar{\mathcal{F}}(t_\Psi, t_\eta) = \mathcal{F}(\text{ortho}(\Psi, D_\Psi, t_\Psi), \eta + t_\eta D_\eta)$;
if $\zeta(t_\Psi, t_\eta) < \nu$ then
    set
        $t_\Psi = \min\left(-\frac{1}{c_1}\frac{\partial \bar{\mathcal{F}}}{\partial t_\Psi}(0,0), \frac{\theta_\Psi}{\|D_\Psi\|}\right), \; t_\eta = \min\left(-\frac{1}{c_2}\frac{\partial \bar{\mathcal{F}}}{\partial t_\eta}(0,0), \frac{\theta_\eta}{\|D_\eta\|_{sF}}\right);$
end if
if $\frac{t_\eta}{t_\Psi} < \underline{c}$ then
    set $t_\Psi = \frac{1}{\underline{c}} t_\eta$, $t_\eta = t_\eta$;
else if $\frac{t_\eta}{t_\Psi} > \bar{c}$ then
    set $t_\Psi = t_\Psi$, $t_\eta = \bar{c} t_\Psi$;
end if
Return $(t_\Psi, t_\eta)$.
```

where $t_\Psi^{\text{trial,min}}, t_\eta^{\text{trial,min}} > 0$, and $\theta_\Psi^{(n)}, \theta_\eta^{(n)} \in (0,1)$ are given parameters. Then we use partial derivatives of $\bar{\mathcal{F}}_n(t_\Psi, t_\eta)$ at $(t_\Psi^{\text{trial}}, t_\eta^{\text{trial}})$ to get $c_{n,1}$ and $c_{n,2}$. Namely, we compute $c_{n,1}$ and $c_{n,2}$ such that $\tilde{\mathcal{F}}_n(t_\Psi, t_\eta) := \bar{\mathcal{F}}_n(0,0) + t_\Psi \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0) + t_\eta \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\eta}(0,0) + \frac{1}{2}c_{n,1}t_\Psi^2 + \frac{1}{2}c_{n,2}t_\eta^2$ satisfies

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

Then we have

$$c_{n,1} = \frac{\frac{\partial \bar{\mathcal{F}}_n}{\partial t_\Psi}(t_\Psi^{\text{trial}}, t_\eta^{\text{trial}}) - \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0)}{t_\Psi^{\text{trial}}}, \quad c_{n,2} = \frac{\frac{\partial \bar{\mathcal{F}}_n}{\partial t_\eta}(t_\Psi^{\text{trial}}, t_\eta^{\text{trial}}) - \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\eta}(0,0)}{t_\eta^{\text{trial}}}. \tag{4.11}$$

We mention that $c_{n,1}$ or $c_{n,2}$ obtained by (4.11) may be less than or equal to 0. In this case, we will use the trial step size (4.10) as the step size in practice. Note that the approximation (4.6) omits the off-diagonal part of the Hessian. It is still open whether or not this approximation is the best, but it does work well when (4.11) is applied as shown in our numerical experiments.

We see that it is needed to calculate the following two partial derivatives:

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

A direct calculation shows

$$\begin{aligned}&\frac{\partial \bar{\mathcal{F}}_n}{\partial t_\Psi}(t_\Psi^{\mathrm{trial}}, t_\eta^{\mathrm{trial}})\\ &\quad= \left\langle \mathcal{F}_\Psi(\mathrm{ortho}(\Psi^{(n)}, D_\Psi^{(n)}, t_\Psi^{\mathrm{trial}}), \eta^{(n)} + t_\eta^{\mathrm{trial}} D_\eta^{(n)}), \frac{\partial\, \mathrm{ortho}(\Psi^{(n)}, D_\Psi^{(n)}, t_\Psi^{\mathrm{trial}})}{\partial t} \right\rangle,\\ &\frac{\partial \bar{\mathcal{F}}_n}{\partial t_\eta}(t_\Psi^{\mathrm{trial}}, t_\eta^{\mathrm{trial}}) = \left\langle \nabla_\eta \mathcal{F}(\mathrm{ortho}(\Psi^{(n)}, D_\Psi^{(n)}, t_\Psi^{\mathrm{trial}}), \eta^{(n)} + t_\eta^{\mathrm{trial}} D_\eta^{(n)}), D_\eta^{(n)} \right\rangle.\end{aligned}$$

It is very difficult to calculate $\frac{\partial\, \mathrm{ortho}(\Psi^{(n)}, D_\Psi^{(n)}, t_\Psi^{\mathrm{trial}})}{\partial t}$. However, it is easy to calculate the first, the second, and the third derivatives of $\mathrm{ortho}(\Psi^{(n)}, D_\Psi^{(n)}, t)$ at $t = 0$. Hence, in practice, we apply a second order approximation:

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

Remark 4.1. In this paper, we focus on the two-parameter step size strategy. In fact, we have also tested two other approximation strategies that use the same step size for Ψ and η. The numerical results show that applying different step sizes for Ψ and η is more efficient than applying the same step size for Ψ and η. The detailed discussions and numerical experiments are provided in sections SM3.1 and SM4.2 in the supplement.

4.2. The preconditioned conjugate gradient method.

Now we introduce the preconditioned conjugate gradient (PCG) method for solving the minimization problem (4.1).

We first introduce the preconditioner applied to $\nabla_\Psi \mathcal{F}$ and $\nabla_\eta \mathcal{F}$. Let $(\Psi, \eta) \in \mathcal{M}^N_{\mathcal{B},\mathbb{C},N_G} \times \mathcal{D}^{N\times N}$. We consider a preconditioner M_Ψ^η for $\nabla_\Psi \mathcal{F}$ as

$$M_\Psi^\eta(\Phi) = M_\Psi\left(\frac{1}{4}\Phi F_\eta^{-1}\right) \quad \forall \Phi \in V_{N_G},$$

where $M_\Psi : V_{N_G} \to V_{N_G}$ is a linear operator. There are some existing works on how to construct an efficient preconditioner M_Ψ, including the following one:

$$[M_\Psi]_{\mathrm{G},\mathrm{G}'} = \delta_{\mathrm{G},\mathrm{G}'}\frac{1}{1 + \frac{1}{2}|\mathrm{k}_0 + \mathrm{G}|^2 + \sqrt{1 + \left(\frac{1}{2}|\mathrm{k}_0 + \mathrm{G}|^2 - 1\right)^2}},$$

which is provided in Quantum ESPRESSO [35] and will be applied in our numerical experiments. We then describe the role of F_η^{-1} in M_Ψ^η. We observe that

$$(\nabla_\Psi \mathcal{F}(\Psi, \eta))_i = 4(H(\rho)\psi_i - (\mathcal{B}\Psi\Sigma)_i)(F_\eta)_{ii}$$

approaches 0 when the occupation number $(F_\eta)_{ii}$ goes to 0. Thus, if we use $(\nabla_\Psi \mathcal{F}(\Psi, \eta))_i$, which corresponds to the small occupation number to update ψ_i, ψ_i

is barely updated. Thus we apply F_η^{-1} to make the norm of each component of $M_\Psi^\eta(\nabla_\Psi \mathcal{F}(\Psi,\eta))$ consistent, by which all Kohn–Sham orbitals can be sufficiently updated. Therefore, the use of F_η^{-1} will improve the rate of the convergence. We refer the reader to [21, 28] for more illustrations of the effect of F_η^{-1} as a preconditioner.

We consider a preconditioner $M_\eta : \mathcal{S}_{\mathbb{C}}^{N\times N} \to \mathcal{S}_{\mathbb{C}}^{N\times N}$ for $\nabla_\eta \mathcal{F}$ as

$$(M_\eta(A))_{ij} = -A_{ij}\frac{1}{2\chi_{ij}} \quad \forall i,j=1,2,\ldots,N, \quad \forall A \in \mathcal{S}_{\mathbb{C}}^{N\times N},$$

where χ_{ij} is defined by (2.6). We see that $M_\eta(\nabla_\eta \mathcal{F}(\Psi,\eta)) = cI_N + \eta - \Sigma$, where $c = \frac{d_\mu}{2\sum_{i'=1}^N \chi_{i'i'}}$ and d_μ is defined by (2.7). Recall $\mathcal{F}(\Psi,\eta) = \mathcal{F}(\Psi,\eta + cI_N)$. Let $\tilde{\eta} = cI_N + \eta$; it is not difficult to see that $M_\eta(\nabla_\eta \mathcal{F}(\Psi,\eta)) = \tilde{\eta} - \Sigma = M_{\tilde{\eta}}(\nabla_{\tilde{\eta}} \mathcal{F}(\Psi,\tilde{\eta}))$ from $\mu(\eta + cI_N) = \mu(\eta) + c$ and Theorem 3.2. We note that $\kappa(\eta - \Sigma)$ is exactly the preconditioned gradient mentioned in [14], where κ is some positive constant and is called a scalar preconditioner.

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 [32, 33], the Fletcher–Reeves (FR) formula [13], the Dai–Yuan (DY) formula [11], etc. In our numerical experiments, we choose the DY formula, which is expressed as

$$(4.12) \qquad \beta^{(n)} = \frac{\mathrm{Re}\left(\left\langle M_{\Psi^{(n)}}^{\eta^{(n)}}(G_\Psi^{(n)}), G_\Psi^{(n)}\right\rangle + \left\langle M_{\eta^{(n)}}(G_\eta^{(n)}), G_\eta^{(n)}\right\rangle\right)}{\mathrm{Re}\left(\left\langle D_\Psi^{(n-1)}, G_\Psi^{(n)} - G_\Psi^{(n-1)}\right\rangle + \left\langle D_\eta^{(n-1)}, G_\eta^{(n)} - G_\eta^{(n-1)}\right\rangle\right)},$$

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 formulae.

We now propose our preconditioned conjugate gradient method as Algorithm 4.2, in which we apply the adaptive two-parameter step size strategy to get the step size.

We see that $D_\Psi^{(n)}$ in the 3rd step of Algorithm 4.2 is not in the tangent space $\mathcal{T}_{\Psi^{(n)}}\mathcal{M}_{\mathcal{B},\mathbb{C},N_G}^N$ and we also expect

$$(4.13) \qquad \frac{\partial \bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0) = \mathrm{Re}\langle \nabla_\Psi \mathcal{F}(\Psi^{(n)},\eta^{(n)}), D_\Psi^{(n)}\rangle.$$

Thus we project $D_\Psi^{(n)}$ onto $\mathcal{T}_{\Psi^{(n)}}\mathcal{M}_{\mathcal{B},\mathbb{C},N_G}^N$ in the 4th step by the projection $P_{\Psi^{(n)}}^*$ to ensure that (4.13) holds. The 5th step is to make sure that search directions are descent directions, i.e., (4.3) holds true. For the 6th step, the choice of $(\theta_\Psi^{(n)}, \theta_\eta^{(n)})$ in our numerical experiments is shown in section 5. As a general algorithm, we do not specify which way or strategy is used for the 8th–9th steps. In fact, people may have more than one choice for them. For example, for the initial step size, we may choose that used in the former iteration as the initial step size for the current step, which is also what we use in the experiments. For $c_{n,1}$ and $c_{n,2}$, we provide three choices, which are (4.11), (SM3.2), and (SM3.3). We should also mention that the update step 11 is based on the Theorems 3.1 and 3.2.

4.3. The restarted preconditioned conjugate gradient method. We turn to consider the restarted PCG method. In fact, the restarting approach has often been used to address the jamming problem for the conjugate gradient method (see, e.g., [18, 30, 34]). Note that the restarting approach for the conjugate gradient method means to set the conjugate parameter $\beta_n = 0$.

Algorithm 4.2. PCG method.

```
1: Given $\alpha \in [0,1)$, $\nu \in (0,1/2]$, $\bar{c} > 1 > \underline{c} > 0$, $t_\Psi^{\min}, t_\eta^{\min}, E_{\rm cut} > 0$, and choose the
   initial data $\Psi^{(0)} \in \mathcal{M}^N_{\mathcal{B},\mathbb{C},N_G}$ and $\eta^{(0)} \in \mathcal{D}^{N\times N}$. Let $D_\Psi^{(-1)} = 0$, $D_\eta^{(-1)} = 0$, $n = 0$;
2: Calculate the gradient $G_\Psi^{(n)} = \nabla_\Psi \mathcal{F}(\Psi^{(n)}, \eta^{(n)})$, $G_\eta^{(n)} = \nabla_\eta \mathcal{F}(\Psi^{(n)}, \eta^{(n)})$ and
   the preconditioned gradient $\widetilde{G}_\Psi^{(n)} = M_{\Psi^{(n)}}^{\eta^{(n)}}(G_\Psi^{(n)})$, $\widetilde{G}_\eta^{(n)} = M_{\eta^{(n)}}(G_\eta^{(n)})$;
3: Calculate the conjugate gradient parameter $\beta^{(n)}$ and the search direction:
   $D_\Psi^{(n)} = -\widetilde{G}_\Psi^{(n)} + \beta^{(n)} D_\Psi^{(n-1)}$, $D_\eta^{(n)} = -\widetilde{G}_\eta^{(n)} + \beta^{(n)} D_\eta^{(n-1)}$;
4: Project the search direction $D_\Psi^{(n)}$ onto the tangent space $\mathcal{T}_\Psi \mathcal{M}^N_{\mathcal{B},\mathbb{C},N_{\rm G}}$

       $$D_\Psi^{(n)} = P^*_{\Psi^{(n)}}(D_\Psi^{(n)});$$

5: Set $D_\Psi^{(n)} = -D_\Psi^{(n)} \operatorname{sign} \operatorname{Re}\langle G_\Psi^{(n)}, D_\Psi^{(n)} \rangle$, $D_\eta^{(n)} = -D_\eta^{(n)} \operatorname{sign} \operatorname{Re}\langle G_\eta^{(n)}, D_\eta^{(n)} \rangle$;
6: Choose the appropriate parameters $(\theta_\Psi^{(n)}, \theta_\eta^{(n)})$;
7: Calculate $\mathcal{C}_n$ by (4.5);
8: Choose an initial guess for the step sizes $(t_\Psi^{n,\rm initial}, t_\eta^{n,\rm initial})$ by some ways;
9: Get $c_{n,1}, c_{n,2}$ by some strategies (e.g., (4.11)) and calculate $t_\Psi^{(n)}, t_\eta^{(n)}$ by

       $$(t_\Psi^{(n)}, t_\eta^{(n)}) = \operatorname{ATPStep}(\Psi^{(n)}, \eta^{(n)}, D_\Psi^n, D_\eta^{(n)}, t_\Psi^{n,\rm initial}, t_\eta^{n,\rm initial}, t_\Psi^{\min}, t_\eta^{\min},$$
       $$\nu, \underline{c}, \bar{c}, c_{n,1}, c_{n,2}, \theta_\Psi^{(n)}, \theta_\eta^{(n)}, \mathcal{C}_n);$$

10: Set $\Psi^{(n+1)} = \operatorname{ortho}(\Psi^{(n)}, D_\Psi^{(n)}, t_\Psi^{(n)})$, $\eta^{(n+1)} = \eta^{(n)} + t_\eta^{(n)} D_\eta^{(n)}$;
11: Pick out $P^{(n+1)} \in \mathcal{O}_{\mathbb{C}}^{N\times N}$ such that $(P^{(n+1)})^* \eta^{(n+1)} P^{(n+1)}$ is diagonal and
    then update

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

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

In practice, we do not expect that search directions are almost orthogonal to the gradients, along which the energy barely changes. Thus, we expect that there exists a positive constant a such that

$$\varlimsup_{n\to\infty} \frac{-\operatorname{Re}\left(\left\langle G_\Psi^{(n)}, D_\Psi^{(n)} \right\rangle + \left\langle G_\eta^{(n)}, D_\eta^{(n)} \right\rangle\right)}{\left|\left\langle G_\Psi^{(n)}, M_{\Psi^{(n)}}^{\eta^{(n)}}(G_\Psi^{(n)}) \right\rangle\right|^a + \left|\left\langle G_\eta^{(n)}, M_{\eta^{(n)}}(G_\eta^{(n)}) \right\rangle\right|^a} > 0. \tag{4.14}$$

Therefore, we reset the conjugate parameter to be 0 when

$$\frac{-\operatorname{Re}\left(\left\langle G_\Psi^{(n)}, D_\Psi^{(n)} \right\rangle + \left\langle G_\eta^{(n)}, D_\eta^{(n)} \right\rangle\right)}{\left|\left\langle G_\Psi^{(n)}, M_{\Psi^{(n)}}^{\eta^{(n)}}(G_\Psi^{(n)}) \right\rangle\right|^a + \left|\left\langle G_\eta^{(n)}, M_{\eta^{(n)}}(G_\eta^{(n)}) \right\rangle\right|^a} < \gamma \tag{4.15}$$

for some given parameter $\gamma \in (0,1)$. Applying this strategy, we obtain a restarted PCG method (denoted by restarted PCG method I), which is essentially Algorithm 4.2 with $\gamma \in (0,1), a > 0$ being added to the 1st step and the following steps being inserted between the 5th and 6th steps:

```
if (4.15) holds then
    $D_\Psi^{(n)} = -P^*_{\Psi^{(n)}}(\widetilde{G}_\Psi^{(n)})$, $D_\eta^{(n)} = -\widetilde{G}_\eta^{(n)}$;
end if
```

In the numerical experiments, we observe that restarting directly is sometimes better than changing the sign of the search direction when the PCG direction is not a descent direction. By this observation, we obtain a new restarted PCG method (denoted by restarted PCG method II), which is exactly Algorithm 4.2 with $\gamma \in (0,1), a > 0$ being added to the1st step and the 5th step being replaced by the following:

```
if $\mathrm{Re}\langle G_\Psi^{(n)}, D_\Psi^{(n)}\rangle \geq 0$ or $\mathrm{Re}\langle G_\eta^{(n)}, D_\eta^{(n)}\rangle \geq 0$ or (4.15) holds then
    $D_\Psi^{(n)} = -P^*_{\Psi^{(n)}}(\widetilde{G}_\Psi^{(n)})$, $D_\eta^{(n)} = -\widetilde{G}_\eta^{(n)}$;
end if
```

We mention that the details of the PCG method and the restarted PCG methods I and II for the case of the multiple k-points are provided in section SM3.2 of the supplement.

4.4. Convergence analysis. In this subsection, we analyze the convergence of the restarted PCG methods. We mention that the convergence of the restarted PCG method for the general sampling $\mathcal{K}$ can be obtained by similar arguments.

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

The following assumption is imposed on the unitarity preserving strategy, which is valid for both QR and polar decomposition (PD) strategies (see, e.g., [9]).

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

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

for any $\Phi \in \mathcal{M}^N_{\mathcal{B},\mathbb{C}}$ and $D_\Phi \in \mathcal{T}_\Phi \mathcal{M}^N_{\mathcal{B},\mathbb{C}}$.

We assume that the preconditioners are positive definite.

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

$$\begin{aligned} \langle \nabla_\Psi \mathcal{F}(\Psi,\eta), M_\Psi^\eta(\nabla_\Psi \mathcal{F}(\Psi,\eta))\rangle &\geq \alpha_\Psi \|\nabla_\Psi \mathcal{F}(\Psi,\eta)\|^2, \\ \langle \nabla_\eta \mathcal{F}(\Psi,\eta), M_\eta(\nabla_\eta \mathcal{F}(\Psi,\eta))\rangle &\geq \alpha_\eta \|\nabla_\eta \mathcal{F}(\Psi,\eta)\|^2_{sF} \end{aligned} \tag{4.16}$$

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

We see that the preconditioner M_Ψ^η we use always satisfies (4.16), and the M_η we apply satisfies (4.16) when f is strictly monotonically decreasing.

The following assumption is imposed on the objective functional.

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

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

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

The following two assumptions are imposed on the parameters $c_{n,1}$ and $c_{n,2}$.

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

$$c_{n,1}+c_{n,2}\le\bar{C}(\|D^{(n)}_\Psi\|^2+\|D^{(n)}_\eta\|^2_{sF}),\quad n=0,1,2,\dots. \tag{4.17}$$

Assumption 4.6. There exist constants $\bar{c}>1>\underline{c}>0$ such that

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

Assumption 4.5 can be viewed as that the Hessian of $\mathcal{F}$ is bounded. If the same step sizes for Ψ and η are used, which is a special case of the adaptive two-parameter step size strategy, then 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.

The following assumption means that some special subsequences of search directions are bounded.

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

$$\lim_{j\to\infty}\frac{-\operatorname{Re}\left(\left\langle G^{(n_j)}_\Psi,D^{(n_j)}_\Psi\right\rangle+\left\langle G^{(n_j)}_\eta,D^{(n_j)}_\eta\right\rangle\right)}{\left|\left\langle G^{(n_j)}_\Psi,M^{\eta^{(n_j)}}_{\Psi^{(n_j)}}(G^{(n_j)}_\Psi)\right\rangle\right|^a+\left|\left\langle G^{(n_j)}_\eta,M_{\eta^{(n_j)}}(G^{(n_j)}_\eta)\right\rangle\right|^a}\ne0$$

with some constant $a>0$, there exists a constant $C>0$ such that

$$\|D^{(n_j)}_\Psi\|+\|D^{(n_j)}_\eta\|_{sF}\le C\quad\forall j\in\mathbb{N}. \tag{4.18}$$

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 also need the following assumption for the step sizes:

$$\varliminf_{n\to\infty}t^{(n)}_\Psi>0,\quad\varliminf_{n\to\infty}t^{(n)}_\eta>0. \tag{4.19}$$

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* $\{(D^{(n)}_\Psi,D^{(n)}_\eta)\}_{n\in\mathbb{N}}$ *satisfies* (4.3) *and* (4.14), *and* $\{(t^{(n)}_\Psi,t^{(n)}_\eta)\}_{n\in\mathbb{N}}$ *satisfies* (4.4) *and* (4.19), *then either* $\|\nabla\mathcal{F}(\Psi^{(n)},\eta^{(n)})\|=0$ *for some positive* n *or*

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

Outline of the proof. We first obtain from (4.4) that

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

which together with (4.19) leads to $\lim_{n\to\infty}\frac{\partial\bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0)=0,\lim_{n\to\infty}\frac{\partial\bar{\mathcal{F}}_n}{\partial t_\eta}(0,0)=0$. Then (4.14) and the expressions of $\frac{\partial\bar{\mathcal{F}}_n}{\partial t_\Psi}(0,0)$ and $\frac{\partial\bar{\mathcal{F}}_n}{\partial t_\eta}(0,0)$ yield the conclusion. ◻

In fact, Lemma 4.8 holds true for more general algorithms—not just for Algorithm 4.2. The details and detailed proof of Lemma 4.8 are provided in section SM3.3.2 of the supplement.

THEOREM 4.9. *Suppose $\mathcal{F}_\Psi$ is bounded, i.e., there exists $C_0 > 0$ such that*

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

and Assumptions 4.2–4.6 *hold true. Let $\{(\Psi^{(n)},\eta^{(n)})\}_{n\in\mathbb{N}}$ be generated by the restarted PCG method I or II. 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 $\|\nabla\mathcal{F}(\Psi^{(n)},\eta^{(n)})\| = 0$ for some $n > 0$ or*

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

Outline of the proof. We first construct $(\theta_\Psi^{(n)},\theta_\eta^{(n)})$ such that

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

As shown in Remark SM3.9 of the supplement, we only need to take into account the subsequence $\{n_j\}_{j\in\mathbb{N}}$ satisfying

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

We also observe that the corresponding $t_\Psi^{(n_j)}$ can only be one of the four types

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

Consequently, there exists a subsequence of $\{n_j\}_{j\in\mathbb{N}}$, which is still denoted by $\{n_j\}_{j\in\mathbb{N}}$ for convenience, such that the step sizes $t_\Psi^{(n_j)} (j = 1,2,\ldots)$ are of the same type (one of the above types). We discuss the four cases of the step sizes $\{t_\Psi^{(n_j)}\}_{j\in\mathbb{N}}$ one by one and then obtain our conclusion with the help of Lemma 4.8. ∎

The detailed proof of Theorem 4.9 is provided in section SM3.3.3 of the supplement. It should be pointed out that, in Theorem 4.9, only the convergence for the gradients of the energy functional is proved. Although a number of numerical experiments show that $\Psi^{(n)}$ and $\eta^{(n)}$ also converge, we are not able to prove it in theory currently.

5. Numerical experiments.

In this section, we apply the PCG method and its restarted versions, which are implemented in software package Quantum ESPRESSO (version 6.4.1) [35], to simulate several gold clusters and two complicated multicomponent periodic systems (see Figures SM1 and SM2 in the supplement for their configurations). All calculations were 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 $\bar{c}$; that is, we do not apply steps 6–10 in Algorithm 4.1. Our algorithms still work well in practice, which shows that the step sizes can be more relaxed in practice, although (4.9) is necessary in our theoretical analysis. Therefore, we directly apply the step sizes (4.8) with $c_{n,1}, c_{n,2}$ being obtained by (4.11) in the numerical simulations. As mentioned below (4.11), we will use the trial step size (4.10) as our step size rather than getting the step size by (4.8) if $c_{n,1} \le 0$ or $c_{n,2} \le 0$.

We observe from our numerical experiments that the choice of $(\theta_\Psi^{(n)}, \theta_\eta^{(n)})$ does not affect the PCG method too much. Thus, we simply choose $(\theta_\Psi^{(n)}, \theta_\eta^{(n)})$ by $\theta_\Psi^{(n)} = \min\{0.8, \|D_\Psi^{(n)}\|\}$ and $\theta_\eta^{(n)} = \min\{0.8, \|D_\eta^{(n)}\|_{sF}\}$. For other cases, we may need to pay more attention to the choice of $(\theta_\Psi^{(n)}, \theta_\eta^{(n)})$. Anyway, it is difficult and open to choose the best $(\theta_\Psi^{(n)}, \theta_\eta^{(n)})$.

In the following tables, we denote the PCG method, the restarted PCG method I, and the restarted PCG method II by PCG, rPCG1, and rPCG2, respectively.

We will compare our PCG method with the SCF iteration approach (see Algorithm SM1.1 in the supplement) for solving the ensemble Kohn–Sham DFT model. For the SCF iteration approach, we have to solve a linear eigenvalue problem at each iteration. In Quantum ESPRESSO, both the Davidson iterative diagonalization and the CG diagonalization are provided to solve the linear eigenvalue problem. Generally speaking, the Davidson iterative diagonalization is faster, but the CG diagonalization uses less memory and is more robust [35]. We shall denote the SCF iterations with the Davidson iterative diagonalization and the CG diagonalization by SCF-David and SCF-CG, respectively.

We now list the parameters used in our numerical experiments. We use the DY approach (4.12) to get the CG parameter and the QR strategy as (4.2) for the orthogonalization operation. We set $t_\Psi^{\text{trial,min}} = t_\eta^{\text{trial,min}} = 0.001$ and initial trial step sizes $t_\Psi^{\text{trial}} = t_\eta^{\text{trial}} = 0.4$. For the restarted versions, we set $\gamma = 0.5$ and $a = 1$. We observed that it holds that $\mathcal{F} - \mathcal{F}_{\min} \propto \|\nabla\mathcal{F}\|^2$ locally from our numerical tests (see Figure SM3 for an example). Therefore, to make the comparison as fair as we can, we set $\|\nabla\mathcal{F}\| < 1.0\times 10^{-5}$ as the convergence criterion for the PCG method and its restarted versions, and the estimated energy error $< 1.0\times 10^{-10}$ as the convergence criterion for the SCF iterations. For the SCF iterations, we choose the Broyden mixing method.

We first see the results for gold clusters. We see that whether or not to restart has almost no effect for the simulations of gold clusters. As a result, we only show the numerical results obtained by the PCG method (Algorithm 4.2) for gold clusters. We also compare the step size strategy applied here with the step size strategies that use the same step size for Ψ and η, which shows that applying different step sizes for Ψ and η is really efficient. The detailed numerical results are provided in section SM4.2 of the supplement.

We compare the PCG method with the SCF iterations for gold clusters in Table 1, where "Iter." means the number of iterations required to terminate the algorithm and "W.C.T." is the total wall clock time spent to converge. In this test, we apply $\sigma = 0.05$ Ry, the Ultrasoft pseudopotentials, and the Gaussian smearing. The density mixing factor for the SCF iterations is 0.3, and the initial guess for the Kohn–Sham orbitals is generated by the superposition of atomic orbitals [35]. Since one step in the SCF iterations is totally different from that in the PCG method, here we do not use the number of iterations to judge the performance of the methods but use the CPU time.

From Table 1 we see that all three methods can obtain convergent approximations for all of the systems. We can also see that the CPU time cost of the PCG method is less than that of SCF-CG but more than that of SCF-David. However, we can observe that as the size of the system increases, the advantage in the CPU time cost by SCF-David over the PCG method becomes less and less obvious, while the advantage of the PCG method over SCF-CG becomes more and more obvious. Besides, we also note that the PCG method always obtains numerical approximations with a smaller

Table 1
Comparison of the SCF iterations and the PCG method.

Algorithm	Energy (Ry)	Iter.	$\|\nabla\mathcal{F}\|$	W.C.T. (s)
Au_{32}	$N_G = 429409$	$N = 211$	$cores = 36$	
SCF-David	−2731.11762824	15	3.5E-05	284.2
SCF-CG	−2731.11762824	19	4.5E-05	1761.2
PCG	−2731.11762824	40	8.6E-06	1095.0
Au_{50}	$N_G = 429409$	$N = 330$	$cores = 36$	
SCF-David	−4267.69535810	17	4.5E-05	557.6
SCF-CG	−4267.69535810	18	5.6E-05	3920.9
PCG	−4267.69535810	41	9.6E-06	1895.7
Au_{92}	$N_G = 556667$	$N = 607$	$cores = 36$	
SCF-David	−7853.07110320	21	5.8E-05	2144.4
SCF-CG	−7853.07110317	23	1.2E-04	17550.0
PCG	−7853.07110320	43	7.4E-06	6043.8
Au_{147}	$N_G = 1320073$	$N = 971$	$cores = 72$	
SCF-David	−12547.62980551	28	1.4E-04	5647.0
SCF-CG	−12547.62980551	31	9.4E-05	40457.7
PCG	−12547.62980551	45	9.2E-06	12413.3
Au_{309}	$N_G = 1320073$	$N = 2040$	$cores = 72$	
SCF-David	−26379.41930504	23	8.1E-05	28833.2
SCF-CG	26379.41930503	25	8.1E-05	160435.7
PCG	−26379.41930507	52	9.4E-06	52807.5

residual $\|\nabla\mathcal{F}\|$ than those obtained by both SCF-CG and SCF-David for these gold clusters, which means that both SCF iteration methods require a smaller convergence threshold to obtain a residual similar to that obtained by the PCG method.

We then see the numerical results for the two complicated periodic systems. The detailed results are reported in Tables 2 and 3. In this test, $\sigma = 0.01$ Ry, the density mixing factor for the SCF iterations is 0.4, and the maximum number of iterations for the SCF is 500. Different from the gold clusters, for these two systems, the spin polarization is taken into account, and the cases using different initial guesses of Kohn–Sham orbitals 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 two typical smearing functions used in the simulation of metallic systems. In Tables 2 and 3, $N_{\mathrm{k}} = 2|\mathcal{K}|$, "atomic" means that the initial guess of Kohn–Sham orbitals is generated by the superposition of atomic orbitals, and "atomic+random" means that the initial guess of Kohn–Sham orbitals is generated by the superposition of atomic orbitals plus a superimposed "randomization" of atomic orbitals [35].

We first see the results obtained by SCF-David and SCF-CG. From Tables 2 and 3, we see that both these two methods fail to converge for the system AlCrTiV within the given maximum number of iterations. However, for the system $\mathrm{NdCu_2Si_2}$, SCF-David converges only for the case with the initial guess of the Kohn–Sham orbitals being given by "atomic," while SCF-CG converges for both cases with different types of the initial guess for the Kohn–Sham orbitals. This validates that SCF-CG is more robust than SCF-David. Comparing the results obtained by our PCG method and restarted PCG methods with the SCF iterations, we see from Tables 2 and 3 that both the PCG method and the restarted PCG methods can obtain convergent approximations for the two systems, no matter which kind of initial guesses and smearing methods are used.

Table 2

Comparison of the SCF iterations, the PCG method, and the restarted PCG methods. The Gaussian smearing is applied.

Algorithm	Initial orbitals	Energy (Ry)	Iter.	$\|\nabla \mathcal{F}\|$
$\mathrm{NdCu_2Si_2}$	$N_G = 3837$	$N = 36$ $N_{\mathrm{k}} = 576$	$cores = 36$	
SCF-David	atomic	−1368.00296219	25	1.2E-05
	atomic+random	−1367.99467076	500	2.7E-03
SCF-CG	atomic	−1368.00296145	44	1.2E-05
	atomic+random	−1368.00296218	51	6.0E-06
PCG	atomic	−1368.00296213	464	9.9E-06
	atomic+random	−1367.99713429	274	9.6E-06
rPCG1	atomic	−1368.00296213	331	9.9E-06
	atomic+random	−1367.99713429	251	9.7E-06
rPCG2	atomic	−1368.00296212	315	8.6E-06
	atomic+random	−1367.99713429	310	8.8E-06
AlCrTiV	$N_G = 1759$	$N = 25$ $N_{\mathrm{k}} = 144$	$cores = 36$	
SCF-David	atomic	−479.31372455	500	6.5E-03
	atomic+random	−479.31277225	500	3.3E-02
SCF-CG	atomic	−479.23303340	500	1.4E-01
	atomic+random	−479.31295489	500	5.9E-03
PCG	atomic	−479.36755754	206	9.0E-06
	atomic+random	−479.36755754	385	9.7E-06
rPCG1	atomic	−479.36755754	144	9.1E-06
	atomic+random	−479.36755754	244	9.8E-06
rPCG2	atomic	−479.36755753	114	7.7E-06
	atomic+random	−479.36755754	118	8.6E-06

We now compare the efficiency of different PCG methods. We first see the results of the system $\mathrm{NdCu_2Si_2}$. From Tables 2 and 3, we can see that for both the case of using the Gaussian smearing and using "atomic" to get the initial guess for the Kohn–Sham orbitals and the case of using the Marzari-Vanderbilt smearing and using "atomic+random" to get the initial guess for the Kohn–Sham orbitals, the number of iterations needed for the PCG method is far larger than those needed for the restarted PCG methods. While for the other cases, the behavior of these different PCG methods is more or less similar. We then see the results of the system AlCrTiV. We can see clearly from Tables 2 and 3 that the number of iterations needed for both of the restarted PCG methods is much less than that needed for the PCG method. Therefore, the restarting strategy does accelerate the convergence of the PCG method. We then compare the performance of two restarted PCG methods. We see that for the system $\mathrm{NdCu_2Si_2}$, the performance of the method rPCG1 is a little better than that of the method rPCG2, while for the system AlCrTiV, the performance of the method rPCG2 is obviously better than that of the method rPCG1.

From these comparisons, we can 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 DFT from a mathematical viewpoint, in which the pseudo-eigenvalue matrix and the general smearing approach are

Table 3

Comparison of the SCF iterations, the PCG method, and the restarted PCG methods. The Marzari–Vanderbilt smearing is applied.

Algorithm	Initial orbitals	Energy (Ry)	Iter.	$\|\nabla\mathcal{F}\|$
$NdCu_2Si_2$	$N_G = 3837$	$N = 36$	$N_{\mathrm{k}} = 576$	$cores = 36$
SCF-David	atomic	-1368.00214304	25	6.8E-06
	atomic+random	-1367.98970427	500	3.0E-02
SCF-CG	atomic	-1368.00214301	40	8.1E-06
	atomic+random	-1368.00214302	46	1.2E-05
PCG	atomic	-1368.00214302	332	9.9E-06
	atomic+random	-1368.00214304	581	1.0E-05
rPCG1	atomic	-1367.99610068	311	8.0E-06
	atomic+random	-1368.00214304	239	9.6E-06
rPCG2	atomic	-1368.00214302	354	9.9E-06
	atomic+random	-1368.00214304	348	9.1E-06
AlCrTiV	$N_G = 1759$	$N = 25$	$N_{\mathrm{k}} = 144$	$cores = 36$
SCF-David	atomic	-479.31161107	500	5.4E-03
	atomic+random	-479.30711971	500	3.9E-02
SCF-CG	atomic	-479.31142973	500	7.5E-02
	atomic+random	-479.30867015	500	1.6E-02
PCG	atomic	-479.36717223	223	9.5E-06
	atomic+random	-479.36717223	158	9.8E-06
rPCG1	atomic	-479.36717223	110	6.7E-06
	atomic+random	-479.36717223	132	9.2E-06
rPCG2	atomic	-479.36717223	97	6.7E-06
	atomic+random	-479.36717223	118	8.2E-06

involved. We have shown the invariance of the energy functional. We have also obtained the existence of the minimizer of the ensemble Kohn–Sham DFT model under some mild and reasonable assumptions. We have then proposed a preconditioned conjugate gradient method to solve the associated energy minimization problem. In particular, we have presented an adaptive two-parameter 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 for the gradients of the energy functional produced by our methods, which are based on the adaptive two-parameter step size strategy. We have reported a number of numerical experiments which not only verify our theory but also show the superiority of our algorithms 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 fail to converge.

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. The authors would also like to thank the anonymous referees and the associate editor for their helpful comments and suggestions, which improved the presentation of this paper.

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), 18A301.

[3] P. E. Blöchl, *Projector augmented-wave method*, Phys. Rev. B, 50 (1994), pp. 17953–17979.

[4] J. Callaway and N. March, *Density functional methods: Theory and applications*, in Solid State Physics, Solid State Phys. 38, Elsevier, New York, 1984, pp. 135–221.

[5] H. Chen, X. Dai, X. Gong, L. He, and A. Zhou, *Adaptive finite element approximations for Kohn–Sham models*, Multiscale Model. Simul., 12 (2014), pp. 1828–1869, https://doi.org/10.1137/130916096.

[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, https://doi.org/10.1137/16M1072929.

[10] X. Dai, L. Zhang, and A. Zhou, *Adaptive step size strategy for orthogonality constrained line search methods*, arXiv:1906.02883, 2020.

[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, https://doi.org/10.1137/S1052623497318992.

[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), 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] M. J. Gillan, *Calculation of the vacancy formation energy in aluminium*, J. Phys. Condens. Matter, 1 (1989), pp. 689–711.

[17] 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), pp. 1999–2014.

[18] W. W. Hager and H. Zhang, *A survey of nonlinear conjugate gradient methods*, Pacific J. Optim., 2 (2006), pp. 35–58.

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

[23] 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.

[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, https://doi.org/10.1137/120880604.

[25] R. M. Martin, *Electronic Structure: Basic Theory and Practical Methods*, 2nd ed., Cambridge University Press, Cambridge, UK; New York, NY, 2020.

[26] N. Marzari. *Ab-initio Molecular Dynamics for Metallic Systems,* Ph.D. 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] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed., Springer, New York, 2006.

[31] R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules*, International Series of Monographs on Chemistry 16, Oxford University Press, New York, 1994.

[32] E. Polak and G. Ribière, *Note sur la convergence de méthodes de directions conjuguées*, Rev. Francaise Informat Recherche Opertionelle, 16 (1969), pp. 35–43.

[33] B. Polyak, *The conjugate gradient method in extremal problems*, USSR Comp. Math. Math. Phys., 9 (1969), pp. 94–112.

[34] M. J. D. Powell, *Restart procedures for the conjugate gradient method*, Math. Program., 12 (1977), pp. 241–254.

[35] *Quantum ESPRESSO*, https://www.quantum-espresso.org/.

[36] R. Remmert, *Theory of Complex Functions*, Grad. Texts in Math. 122, Springer-Verlag, New York, 1991.

[37] 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), 054107.

[38] 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.

[39] N. Troullier and J. L. Martins, *Efficient pseudopotentials for plane-wave calculations*, Phys. Rev. B, 43 (1991), pp. 1993–2006.

[40] 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, https://doi.org/10.1137/14098973X.

[41] D. Vanderbilt, *Soft self-consistent pseudopotentials in a generalized eigenvalue formalism*, Phys. Rev. B, 41 (1990), pp. 7892–7895.

[42] 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, https://doi.org/10.1137/S1052623403428208.

[43] X. Zhang, J. Zhu, Z. Wen, and A. Zhou, *Gradient type optimization methods for electronic structure calculations*, SIAM J. Sci. Comput., 36 (2014), pp. C265–C289, https://doi.org/10.1137/130932934.

[44] 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, https://doi.org/10.1137/140967994.

[45] 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), 033305.