A PROXIMAL GRADIENT METHOD FOR ENSEMBLE DENSITY FUNCTIONAL THEORY

MICHAEL ULBRICH[†], ZAIWEN WEN[‡], CHAO YANG[§], DENNIS KLÖCKNER[†], AND ZHAOSONG LU[¶]

Abstract. The ensemble density functional theory is valuable for simulations of metallic systems due to the absence of a gap in the spectrum of the Hamiltonian matrices. Although the widely used self-consistent field iteration method can be extended to solve the minimization of the total energy functional with respect to orthogonality constraints, there is no theoretical guarantee on the convergence of these algorithms. In this paper, we consider an equivalent model with a single variable and a single spherical constraint by eliminating the dependence on the fractional occupancies. A proximal gradient method is developed by keeping the entropy term but linearizing all other terms in the total energy functional. Convergence to the stationary point is established. Numerical results in the KSSOLV toolbox under the Matlab environment show that they can outperform SCF consistently on many metallic systems.

Key words. Ensemble Density functional theory, Kohn-Sham total energy minimization, Orthogonality/spherical constraints, Proximal Gradient Method

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

1. Introduction.

Kohn-Sham Density functional theory (KS-DFT) [19, 21] has been widely used in ab initio electronic-structure calculations to accurately predict electronic, optical, structural and other properties of molecules and solids. It is often formulated as a constrained electron total energy minimization problem or a nonlinear eigenvalue problem known as the Kohn-Sham equations. The Kohn-Sham equations are the first-order necessary optimality condition of the constrained energy minimization problem. With the fundamental development of both theory and algorithms for DFT coupled with the increasingly powerful computer resources, it is now possible to simulate many materials properties with a precision that is often comparable to that of real experiments.

It is well known that the standard formulation of the Kohn-Sham problem is more difficult to solve numerically for metals than for insulators or semiconductors [12, 11, 17, 23, 29, 36]. For insulators and semiconductors, there exists a gap in the spectrum of the Kohn-Sham Hamiltonian near the vicinity of what is known as the Fermi level or chemical potential. For metallic systems, such a gap is very small or absent. Note that our definition of metallic systems pertains to both periodic systems (solids) and isolated systems such as molecules and atoms although in this paper, we will focus mainly on isolated systems with no special symmetry in our numerical study. The existence of a gap in the spectrum of the Kohn-Sham Hamiltonian makes invariant subspace associated with eigenvalues below the gap well defined and relatively easy to compute. Without such a gap, the standard formulation of Kohn-Sham problem becomes ill-posed because the desired invariant subspace cannot be easily separated. This ill-posedness leads to some ambiguity in the definition of the electron charge density and numerical difficulty in obtaining a self-consistent solution.

To overcome the difficulty of the standard formulation for metallic systems, alternative formulations have been developed. These formulation include the finite-temperature statistical ensembles [30] and the finite-temperature ensemble DFT (E-DFT) [11, 17, 23, 29, 36]. In these formulations, each KS eigenpair is associated with an occupation

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

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

[4] Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, UNITED STATES (cyang@lbl.gov). Support for this work was provided through the Scientific Discovery through Advanced Computing (SciDAC) program funded by U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research (and Basic Energy Sciences) under award number DE-SC0008666.

[5] Department of Mathematics, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada. (zhaosong@sfu.ca). Research supported in part by NSERC Discovery Grant.

number that assumes a value between 0 and 1. The total energy of the electronic system contains an additional temperature dependent entropy term that reflects the statistical nature of electron occupation. Other extensions of the standard KSDFT model include the extended KSDFT models defined in [1, 10, 16] which we will discuss in section 2.3. Numerical methods have been developed to compute the solution of these models.

Broadly speaking, there are two classes of methods for computing the solution to the standard Kohn-Sham problem. The first class of methods focus on the first order necessary condition, and use a Newton-like method to find the approximate solution of the Kohn-Sham problem. This class of methods is often called a self-consistent field iteration (SCF). The interested reader is referred to [5, 6, 7, 8, 9] for discussions of the SCF iteration and its theoretical properties. The second class of methods try to minimize the total energy functional directly [2, 4, 32, 33, 39, 40, 25, 31]. They are sometimes called direct minimization methods. The first class of methods have been extended to compute solutions of the ensemble Kohn-Sham DFT problem [11, 17, 29, 36]. In this approach, the Fermi-level is introduced as an extra degree of freedom that is solved iteratively. The occupation numbers can be readily evaluated once the eigenvalues of the KS Hamiltonian and the Fermi-level are known. There has not been a significant amount of effort on extending direct minimization methods to solve the ensemble DFT problem. We will focus on this approach in this paper.

One of the challenges with developing a direct minimization method for EDFT is the need to treat occupation numbers as additional degrees of freedom to be optimized. These extra degrees of freedom makes modification of the existing algorithm nontrivial. We will show that an explicit parameterization of the occupation number can be avoided by an equivalent alternative formulation of the EDFT minimization problem.

One of the major difficulties in direct minimization methods is related to orthogonality constraints. A direct constrained minimization (DCM) algorithm is designed in [48] where the new search direction is built from a subspace spanned by the current approximation to the optimal wavefunction, the preconditioned gradient and the previous search direction. Trust region methods [14, 15, 37, 38, 47] substitute the linear eigenvalue problem in SCF by the so called trust-region subproblems and monotonic reduction of the total energy can be achieved by imposing a suitable update of the trust region radius. The adaptive regularized SCF approach in [41] takes advantage of the Hessian of the total energy functional and establishes rigorous global convergence to the first-order optimality conditions. A projected gradient-type method is studied in [42, 50]. One computational benefit is that the computation of linear eigenvalue problems is no longer needed. The main costs of the approach arise from the assembling of the total energy functional and its gradient on manifold and the projection onto the manifold. These tasks are cheaper than eigenvalue computation and they are often more suitable for parallelization as long as the evaluation of the total energy functional and its partial derivative is efficient. Numerical results in [50] based on the software package Octopus[1] show that they can outperform SCF consistently on large molecular systems, including carbon nano-tube, biological ligase 2JMO ($C_{178}H_{283}O_{50}N_{57}S$) and protein fasciculin2 ($C_{276}H_{442}O_{90}N_{88}S_{10}$).

In this paper, we aim to develop gradient type approaches for solving the ensemble DFT. Notice that the gradient approaches in [42, 50] can be applied to the ensemble DFT directly due to the additional variable and constraints on the fractional occupancies. We first show a derivation of an alternative model with a single spherical constraint. The variable on the fractional occupancies is eliminated and only one variable, denoted by Z (which is a $n \times p$ complex matrix), is involved in the new model. The entropy term plays a role as a regularization term on the singular value of the variable Z. Our model is related to, but different from the one-body density formulation considered in [1, 10, 16] whose variable is the $n \times n$ density matrix. Since the gradient of the entropy tends to blow up if some of the singular values of Z are close to 0 or 1, our experiences show that the gradient approaches in [42, 50] encountered numerical difficulties as the iteration converges. It is also not obvious to extend the regularized SCF method [41] to solve the

[1] OCTOPUS. http://www.tddft.org/programs/octopus.

ensemble DFT because this approach requires the computation of the Hessian-vector product of objective function but evaluating the Hessian-vector product of the entropy term in our new model is impractical. Inspired by the recent progress on sparse and low-rank matrix optimization [18, 43, 44, 28] for minimizing a summation of a Lipschitz continuously differentiable function and a possibly nonsmooth function, we design a proximal gradient methods for our model. In each subproblem, the entropy term is kept but other terms are linearized and a proximal term is added to ensure convergence. Although the subproblem has no explicit closed-form solution, it can be solved efficiently using a Lagrangian approach. Convergence to a point satisfying the first-order optimality conditions is proved. The numerical performance of our proximal gradient methods is further improved by the state-of-the-art acceleration techniques such as Barzilai-Borwein steps and non-monotone line search with global convergence guarantees. Our approaches can quickly reach the vicinity of an optimal solution and produce a moderately accurate approximation, at least in our numerical examples.

This paper is organized as follows. In Section 2, we review the model of the ensemble DFT, derive the first-order optimality conditions and present the corresponding SCF iteration method. An equivalent model with a single spherical constraint is proposed in Section 2.5. A nonmonotone proximal gradient method, the techniques on solving the proximal subproblem and the convergence analysis are presented in Section 3. Finally, we demonstrate the robustness and efficiency of our algorithms based on KSSOLV in Section 4.

1.1. Notation and Preliminaries.

For a matrix $X \in \mathbb{C}^{m\times n}$, the matrices $\overline{X}, X^*, \mathrm{Re}(X)$ and $\mathrm{Im}(X)$ denote the complex conjugate, the complex conjugate transpose, and the real and imaginary parts of X, respectively. The set of $n\times n$ Hermitian matrix is denoted by $\mathcal{S}^{n\times n}$, $I_n \in \mathbb{R}^{n\times n}$ stands for the n-dimensional identity matrix and e denotes the vector of all ones. The notation $X \succeq 0$ means that the matrix $X \in \mathcal{S}^{n\times n}$ is positive semidefinite. For a vector $d \in \mathbb{C}^n$, the operator $\mathrm{Diag}(d)$ returns a square diagonal matrix in $\mathbb{C}^{n\times n}$ with the elements of d on the main diagonal, while conversely $\mathrm{diag}(X)$ returns the vector in $\mathbb{C}^n$ containing the main diagonal elements of the square matrix $X \in \mathbb{C}^{n\times n}$. The trace of X, i.e., the sum of the elements on the main diagonal of a square matrix $X \in \mathbb{C}^{n\times n}$, is denoted by $\mathrm{tr}(X)$. The Frobenius inner product for matrices $X, Y \in \mathbb{C}^{m\times n}$ is defined as $\langle X, Y\rangle = \mathrm{tr}(X^*Y)$ and the corresponding Frobenius norm $\|\cdot\|_F$ is given by $\|X\|_F = \langle X, X\rangle^{1/2} = \big(\sum_{i,j} |X_{ij}|^2\big)^{1/2}$. The Hadamard product $X \circ Y$ is defined componentwise as $(X \circ Y)_{ij} = X_{ij}Y_{ij}$. For a Hermitian matrix $F \in \mathbb{C}^{n\times n}$ the operators $\lambda_i(F), \lambda_{\text{min}}(F)$ and $\lambda_{\text{max}}(F)$ denote the i-th, the smallest and the largest eigenvalue, respectively. Given two matrices $X \in \mathbb{C}^{m\times n}$ and $Y \in \mathbb{C}^{n\times m}$, the trace identity is $\mathrm{tr}(XY) = \mathrm{tr}(YX)$. For $X, Y \in \mathbb{C}^{m\times n}$ and $v \in \mathbb{C}^n$, it holds

$$(X \circ Y)v = \mathrm{diag}(X\mathrm{Diag}(v)Y^*). \tag{1.1}$$

2. Preliminaries.

2.1. The KSDFT Model.

The KSDFT model [22] expresses the total energy of an interacting many-electron system in terms of single-electron wavefunctions associated with a non-interacting reference system. When spin is ignored, the KS total total energy functional can be written as [46]

$$E_{KS}(\{\psi_i\}) = \frac{1}{2}\sum_{i=1}^{p_e}\int_\Omega \|\nabla\psi_i(r)\|^2 dr + \int_\Omega \rho(r)V_{ion}(r)dr + \frac{1}{2}\int_\Omega\int_\Omega \frac{\rho(r)\rho(r')}{\|r-r'\|}drdr' + E_{xc}(\rho), \tag{2.1}$$

where p_e represent the total number of (valence) electron pairs, ψ_i, $i = 1, 2, \ldots, p_e$ are single-particle wavefunctions that satisfy the orthonormality constraint $\int \psi_i^*\psi_j = \delta_{ij}$, $r \in \mathbb{R}^3$, $\Omega \subseteq \mathbb{R}^3$ and the charge density $\rho(r)$ is defined by $\rho(r) = \sum_{i=1}^{p_e} |\psi_i(r)|^2$. The function $V_{ion}(r) = \sum_{j=1}^{n_u} q_j/\|r - \hat{r}_j\|$ represents the ionic potential induced by the n_u nuclei, where $\hat{r}_j$ and q_j are the position and charge of the jth nucleus, respectively. $E_{xc}(\rho)$ is the exchange-correlation energy, which accounts for the non-interacting reference system fails to capture. Finding the ground state energy of the system is equivalent to solving the following minimization problem

$$\begin{aligned} &\inf_{\psi_i} \quad E_{KS}(\{\psi_i\}) \\ &\text{s.t.} \int_\Omega \psi_i(r)^* \psi_j(r) \mathrm{d}r = \delta_{ij}, \quad 1 \leqslant i, j \leqslant p_e. \end{aligned} \tag{2.2}$$

The continuous KSDFT model can be discretized by either a planewave expansion scheme or real space approaches including finite difference, finite element, finite volume and wavelet methods. Using a suitable discretization scheme whose spatial degree of freedom is n (for more details we refer the reader to [46]), the electron wavefunctions can be approximated by a matrix $X = [x_1, \ldots, x_{p_e}] \in \mathbb{C}^{n \times p_e}$. They satisfy the orthogonality constraint $X^*X = I$ since the wavefunctions X must be orthogonal to each other due to physical constraints. Here, it is assumed that the basis chosen in the discretization is orthonormal. The discretized charge density can be expressed as

$$\rho(X) = \operatorname{diag}(XX^*) = \sum_{i=1}^{p_e} |x_i|^2. \tag{2.3}$$

To simplify notation, we write the discretized total energy function as:

$$E(X) = \frac{1}{2} \operatorname{tr}(X^* L X) + \operatorname{tr}(X^* V_{ion} X) + \frac{1}{2} \rho^\top L^\dagger \rho + e^\top \epsilon_{xc}(\rho) \tag{2.4}$$

where L is a finite dimensional representation of the Laplacian operator, V_{ion} is the ionic pseudopotentials sampled on a suitably chosen Cartesian grid, $L^\dagger$ corresponds to the pseudo-inverse of L and $\epsilon_{xc}(\rho)$ denotes the exchange correlation energy function evaluated on a spatial grid. Our simplification does not assume a particular type of discretization scheme and thus does not include additional coefficient matrices (e.g., numerical quadrature weight matrices, basis overlap matrices) associated with certain type of discretization schemes. One can think of L as a simple finite difference approximation to the Laplacian operator on a uniform grid, and assume a simple quadrature rule is used to perform all numerical integration. Such simplification does not fundamentally affect the algorithm we describe in this paper or the convergence analysis we present.

Once (2.2) is discretized, we can express the finite dimensional KSDFT model as

$$\min_X \quad E(X) \quad \text{s.t.} \quad X^*X = I. \tag{2.5}$$

The first order necessary condition becomes a nonlinear eigenvalue problem

$$H(X)X = X\Lambda, \quad X^*X = I, \tag{2.6}$$

where $H(X) = L + V_{ion} + \operatorname{Diag}(\operatorname{Re}(L^\dagger)\rho) + \operatorname{Diag}(\mu_{xc}^* e)$ with $\mu_{xc} = \frac{\partial \varepsilon_{xc}}{\partial \rho} \in \mathbb{R}^{n \times n}$, and Λ is a diagonal matrix containing the p_e smallest eigenvalues of $H(X)$.

2.2. Other DFT Models. When the p_eth eigenvalue of $H(X)$ is a multiple (or degenerage) eigenvalue, the charge density ρ in (2.3) is not well defined. Even if there is a gap between the p_eth and p_e+1st eigenvalue, finding the optimal solution to (2.5) or (2.6) can be difficult if the gap is very small. The reason is that numerically it is difficult to separate the invariant subspace associated with the smallest p_e eigenvalues from eigenvector associated with the next few eigenvalues. In that sense, the KSDFT problem can be considered ill-posed. This situation often occurs for metallic systems or molecules that contain transition metal atoms [12, 11, 17, 23, 29, 36]. To overcome numerical difficulties associated with these types of systems, the standard KSDFT model is often modified to avoid defining the charge density by (2.3).

2.3. Extended DFT model. One type of modification involves working with the one-body density matrix instead of X directly. If there is a gap between the p_eth and p_e+1st eigenvalues of H, then the one-body density matrix can be defined as $D = XX^*$. In this case, $\rho(X) = \hat{\rho}(D) := \mathrm{diag}(D)$. Using the simplified notation $A \equiv \frac{1}{2}L + V_{ion}$. we can express the total energy (2.4) as a function of D, i.e.

$$E(X) = \hat{E}(D) := \mathrm{tr}(AD) + \frac{1}{2}\hat{\rho}^\top L^\dagger \hat{\rho} + e^\top \epsilon_{xc}(\hat{\rho}). \tag{2.7}$$

In fact, the constraints $X^*X = I$ can be expressed equivalently in terms of $D \in \mathbb{C}^{n\times n}$ such that

$$D = D^*, \quad \mathrm{tr}(D) = p_e, \quad \text{and} \quad D = D^2,$$

since the constraint $D = D^2$ requires that the eigenvalues of D must be either 0 or 1 and $\mathrm{tr}(D) = \sum_{i=1}^n \lambda_i(D) = p_e$. As a result, the KSDFT model (2.5) can be rewritten in terms of a one-body density matrix [1, 10, 16]:

$$\min_{D\in\mathbb{C}^{n\times n}} \quad \hat{E}(D) \quad \text{s.t.} \quad D = D^*, \quad \mathrm{tr}(D) = p_e, \quad D = D^2. \tag{2.8}$$

By relaxing $D = D^2$ constraint to allow $0 \le \lambda_i(D) \le 1$, one can obtain an extended KSDFT model [1]

$$\min_{D\in\mathbb{C}^{n\times n}} \quad \hat{E}(D) \quad \text{s.t.} \quad D = D^*, \quad \mathrm{tr}(D) = p_e, \quad 0 \le \lambda(D) \le 1. \tag{2.9}$$

A gradient projected algorithm is proposed in [10] to find the solution of (2.9).

2.4. Ensemble DFT model. Another way to modify the KSDFT model is to allow more wavefunctions to be included in the definition of the charge density, i.e.,

$$\rho(r) = \sum_{i=1}^{p} f_i |\psi_i(r)|^2$$

for some $p \ge p_e$, and use fractional occupation $0 \le f_i \le 1$ to ensure that the charge density sums up to the number of electron pairs. This implies that

$$\sum_{i=1}^{p} f_i = p_e.$$

To account for fractional occupation, an entropy terms $R(f)$ scaled by a temperature factor $\alpha = \kappa_B T$, where κ_B is the Boltzmann constant, is introduced in the energy functional. The entropy term has the form $R(f) = \sum_{i=1}^{p} s(f_i)$, where

$$s(t) = \begin{cases} t\ln t + (1-t)\ln(1-t), & 0 < t < 1, \\ 0, & \text{otherwise.} \end{cases} \tag{2.10}$$

This approach yields what is often known as a finite temperature KSDFT [11, 17, 29, 36] or an ensemble KSDFT model (EDFT). In this EDFT model, we are concerned with minimizing the Helmholtz free energy functional

$$\begin{aligned} E_{EDFT}(\{\psi_i\}, f) = \frac{1}{2}\sum_{i=1}^{p}\int_\Omega f_i\|\nabla\psi_i(r)\|^2 dr + \int_\Omega \rho(r)V_{ion}(r)dr + \frac{1}{2}\int_\Omega\int_\Omega \frac{\rho(r)\rho(r')}{\|r-r'\|}drdr' \\ +E_{xc}(\rho) + \alpha R(f). \end{aligned} \tag{2.11}$$

Therefore, in the EDFT model, the minimization problem to be solved is

$$\begin{aligned} \inf_{\psi_i, f} \quad & E_{EDFT}(\{\psi_i\}, f) \\ \text{s.t.} \int_\Omega & \psi_i(r)^*\psi_j(r)\mathrm{d}r = \delta_{ij}, \quad 1 \leqslant i, j \leqslant p, \\ & e^\top f = p_e, \quad 0 \le f \le 1. \end{aligned} \tag{2.12}$$

Similar to the KSDFT, the electron wavefunctions are approximated by $X = [x_1, \dots, x_p] \in \mathbb{C}^{n\times p}$ using a suitable discretization scheme. Then the discretized charge density of electrons in ensemble DFT can be written as

$$\rho(X, f) := \mathrm{diag}(X\mathrm{Diag}(f)X^*). \tag{2.13}$$

It is obvious that $\rho(X, f)$ is real. The corresponding discretized total energy functional (2.11) is

$$M(X, f) = \mathrm{tr}(\mathrm{Diag}(f)X^*AX) + \frac{1}{2}\rho^\top L^\dagger\rho + e^\top\epsilon_{xc}(\rho) + \alpha R(f). \tag{2.14}$$

The discretized ensemble total energy minimization problem becomes

$$\begin{aligned} \min_{X\in\mathbb{C}^{n\times p}, f\in\mathbb{R}^p} \quad & M(X, f) \\ \text{s.t.} \quad & X^*X = I, \\ & e^\top f = p_e \quad 0 \le f \le 1. \end{aligned} \tag{2.15}$$

If $\bar{D} = X\mathrm{Diag}(f)X^*$, then we have $\rho(X, f) = \hat{\rho}(\bar{D}) := \mathrm{diag}(\bar{D})$ and

$$M(X, f) = \hat{M}(\bar{D}) := \mathrm{tr}(A\bar{D}) + \frac{1}{2}\hat{\rho}^\top L^\dagger\hat{\rho} + e^\top\epsilon_{xc}(\hat{\rho}) + \alpha R((\lambda_1(\bar{D}), \dots, \lambda_p(\bar{D}))^\top).$$

Therefore, the one-body density matrix formulation of (2.15) is

$$\min_{\bar{D}\in\mathbb{C}^{n\times n}} \quad \hat{M}(\bar{D}) \quad \text{s.t.} \quad \bar{D} = \bar{D}^*, \quad \mathrm{tr}(\bar{D}) = p_e, \quad 0 \le \lambda(\bar{D}) \le 1, \quad \text{rank}(\bar{D}) \le p. \tag{2.16}$$

The rank constraint is imposed since X is a rank p matrix in (2.15).

The Lagrangian function of (2.15) is

$$\mathcal{L}(X, f, \Lambda, \mu) := M(X, f) - \langle\Lambda, X^*X - I\rangle - \mu(e^\top f - p_e), \tag{2.17}$$

where $\Lambda \in S^{p\times p}$ and μ are the Lagrangian multipliers. The Hamiltonian matrix is

$$H(X,f) = A + \mathrm{Diag}(\mathrm{Re}(L^{\dagger})\rho) + \mathrm{Diag}(\mu_{xc}^{*} e),$$

where $\mu_{xc} = \frac{\partial \varepsilon_{xc}}{\partial \rho} \in \mathbb{R}^{n\times n}$. For any $f \in \mathbb{R}^p$, it follows from the identity (1.1) that

$$\rho(X,f) = \mathrm{diag}(X\mathrm{Diag}(f)X^{*}) = \left(X \circ \overline{X}\right) f.$$

Hence, it can be verified that the first-order optimality conditions of (2.15) is

$$\frac{\partial \mathcal{L}}{\partial f_i} = x_i^{*} H(X,f) x_i + \alpha \ln\left(\frac{f_i}{1-f_i}\right) - \mu = 0, \quad i = 1,\dots,p, \tag{2.18}$$

$$e^{\top} f = p_e, \quad 0 \le f \le 1, \tag{2.19}$$

$$\frac{\partial \mathcal{L}}{\partial X} = 2(H(X,f)X\mathrm{Diag}(f) - X\Lambda) = 0, \tag{2.20}$$

$$X^{*}X = I. \tag{2.21}$$

Equation (2.18) yields

$$f_i = \frac{1}{1+e^{(\gamma_i - \mu)/\alpha}}, \tag{2.22}$$

where $\gamma_i = x_i^{*} H(X,f) x_i$. Substituting (2.22) into (2.19) gives

$$G(\mu) = \sum_{i=1}^{p} \frac{1}{1+e^{(\gamma_i-\mu)/\alpha}} = p_e. \tag{2.23}$$

It is easy to verify that $G(\mu)$ is monotonic with respect to μ. Hence, the solution of (2.23) is unique, and it can be obtained by, for example, using a bisection algorithm.

Equation (2.20) and the orthogonality constraints (2.21) imply that $\Lambda = X^{*}H(X,f)X\mathrm{Diag}(f)$. This equation suggests that a local minimizer of $\mathcal{L}$ is invariant under $H(X,f)$. Although the symmetric Lagrange multiplier Λ may not necessarily be diagonal, an unitary transformation W can always be chosen such that $(XW)^{*}H(X,f)XW = \Omega$ is diagonal. Such an unitary transform allows us to turn Λ into a diagonal matrix.

Therefore, the well-known self-consistent field iteration (SCF) method can be extended to solve problem (2.15). Starting from an initial orthogonal matrix X^0, the Lagrangian multiplier μ^k at k-th iteration is computed by solving (2.23) with $\gamma_i^k = (x_i^k)^{*} H(X^{k-1}, f^{k-1}) x_i^k$. Consequently, f^k is updated with μ^k by using (2.22). Then one solve the linear eigenvalue problem

$$\begin{aligned} H(X^{k-1}, f^k)X &= \quad X\Omega, \\ X^{*}X &= \qquad I. \end{aligned} \tag{2.24}$$

The procedure is iteratively performed until convergence is met. An outline of the SCF method is described in Algorithm 1.

Algorithm 1: A SCF Method for EDFT

```
Given a feasible $X^0$ and $f^0$. Set $k = 0$.
while "convergence" is not met do
    Compute $\mu^k$ by solving (2.23).
    Compute $f^k$ using (2.22).
    Solve the linear eigenvalue problem (2.24) to obtain $X^k$.
    Set $k = k + 1$.
```

Although the SCF iteration with charge density or potential mixing works well on many problems, few theoretical convergence analysis is available. The existing convergence analysis [34, 45, 26, 27] often requires strong assumptions on the gap between the occupied and unoccupied states to guarantee either local or global convergence of the SCF iteration. When the local density approximation (LDA) or generalized gradient approximation (GGA) are used for the exchange-correlation function, the main computational bottleneck of the SCF iteration is in solving a sequence of linear eigenvalue problems. Since the number of degrees of freedom n and the number of occupied states p_e can be very large, the computational cost of the SCF iteration is often dominated by that associated with computing the desire eigenpairs.

2.5. An equivalent EDFT model with a single spherical constraint. Problem (2.15) is difficult to solve due to the unitary constraints $X^*X = I$ and the interaction between the variables X and f. In this section, we present an equivalent model with a single spherical constraint. It is simpler since the unitary constraints are eliminated and only one variable is involved. We should point out that our model is related to, but different from the one-body density matrix formulation in [1, 10, 16] due to the presence of the entropy term. Furthermore, our approach represents $\bar{D} = X\mathrm{Diag}(f)X^*$ in the form $\bar{D} = ZZ^*$ and works with $Z \in \mathbb{C}^{n\times p}$ as the single optimization variable. This formulation is more general than working with the square root $\bar{D}^{1/2}$ of the one-body density matrix and allows us to incorporate the rank constraint $\mathrm{rank}(\bar{D}) \le p$ directly into the column dimension of Z. A further advantage is that the derivative of commonly used exchange correlation energies with respect to $\bar{D}$ is usually not Lipschitz continuous at points where $\hat{\rho}(\bar{D}) = 0$, whereas for $\bar{D} = ZZ^*$ the derivative with respect to Z has this property, see Lemma 3.3.

To introduce the model, let $Z \in \mathbb{C}^{n\times p}$, define the density

$$\rho = \mathrm{diag}(ZZ^*)$$

and the energy functional

$$E(Z) := \mathrm{tr}(Z^*AZ) + \frac{1}{2}\rho^\top L^\dagger \rho + e^\top \epsilon_{xc}(\rho). \tag{2.25}$$

For $Z \in \mathbb{C}^{n\times p}$, let $\sigma_i(Z)$ denote the ith largest singular value of Z for $i = 1, \dots, p$, $\sigma(Z) = (\sigma_1(Z), \dots, \sigma_p(Z))^\top$ and

$$\Omega(Z) = \{(U, V) \in \mathbb{C}^{n\times p} \times \mathbb{C}^{p\times p} : U^*U = V^*V = I, Z = U\mathrm{Diag}(\sigma(Z))V^*\}. \tag{2.26}$$

Then our new ensemble DFT model is

$$
\begin{aligned}
\min_{Z\in\mathbb{C}^{n\times p}} \quad & \mathcal{M}(Z) := E(Z) + \alpha R(Z^*Z) \\
\text{s.t.} \quad & \|Z\|_F = \sqrt{p_e}, \\
& 0 \le \sigma_i(Z) \le 1, \quad i = 1,\dots,p,
\end{aligned}
\tag{2.27}
$$

where the spectral function R is defined by $R(Z^*Z) = R(\sigma(Z)^2)$ with $(\sigma(Z)^2)_i = \sigma_i(Z)^2$.

The next theorem shows that models (2.15) and (2.27) are equivalent.

THEOREM 2.1. *The following statements are true.*

1. *If (X, f) is an optimal solution of* (2.15)*, then $Z = X\mathrm{Diag}(f^{\frac{1}{2}})$ is an optimal solution of* (2.27).
2. *Suppose that Z is an optimal solution of* (2.27)*. Let $Z = X\mathrm{Diag}(w)V^*$ be the SVD of Z such that X and V are orthogonal and w are the singular values. Then (X, w^2) is an optimal solution of* (2.15).

Proof. Let (X, f) be a minimizer of (2.15) and $\bar{Z}$ be an optimal solution of (2.27). Since $0 \le f \le 1$, we can set $Z = X\mathrm{Diag}(f^{\frac{1}{2}})$. It can be verified that $\mathrm{tr}(\mathrm{Diag}(f)X^*AX) = \mathrm{tr}(Z^*AZ)$, $\rho = \mathrm{diag}(X\mathrm{Diag}(f)X^*) = \mathrm{diag}(ZZ^*)$, and $Z^*Z = \mathrm{Diag}(f)$. This shows that $\mathrm{tr}(Z^*Z) = e^\top f = p_e$ and $\{f_i\}$ are the eigenvalues of Z^*Z, i.e., the squared singular values of Z. Hence, Z is feasible for (2.27) and there holds $R(Z^*Z) = R(f)$ and $\mathcal{M}(Z) = M(X, f)$. The optimality of $\bar{Z}$ yields $\mathcal{M}(\bar{Z}) \le \mathcal{M}(Z) = M(X, f)$.

Conversely, let $\bar{Z} = \bar{X}\mathrm{Diag}(\bar{w})\bar{V}^*$ be the SVD of $\bar{Z}$ such that $\bar{X}$ and $\bar{V}$ are orthogonal and $\bar{w}$ are the singular values. Let $\bar{f}_i = \bar{w}_i^2$. Then it can be proved similarly that $(\bar{X}, \bar{f})$ is a feasible solution of (2.15) and $M(\bar{X}, \bar{f}) = \mathcal{M}(\bar{Z})$. The optimality of (X, f) and $\bar{Z}$, respectively, yield

$$M(X, f) \le M(\bar{X}, \bar{f}) = \mathcal{M}(\bar{Z}) \le \mathcal{M}(Z) = M(X, f).$$

Thus, Z is an optimal solution of (2.27) and $(\bar{X}, \bar{f})$ is an optimal solution of (2.15), which completes the proof of statements 1 and 2. □

REMARK 2.2. *Theorem 2.1 is formulated for optimal solutions. In a similar way, one can show that for any feasible Z of* (2.27) *there exists a feasible solution (X, f) of* (2.15) *with $\mathcal{M}(Z) = M(X, f)$ and vice versa.*

Problem (2.27) is invariant under unitary transformations. Specifically, let $U \in \mathbb{C}^{p\times p}$ be unitary and set $\widehat{Z} = ZU$. It holds that Z is a minimizer of (2.27) if and only if $\widehat{Z}$ is a minimizer of (2.27). The energy functional (2.25) is the same as the Kohn-Sham energy functional. Hence, we define the Hamiltonian matrix

$$H(Z) = A + \mathrm{Diag}(\mathrm{Re}(L^\dagger)\rho) + \mathrm{Diag}(\mu_{xc}^* e),$$

where $\mu_{xc} = \frac{\partial \varepsilon_{xc}}{\partial \rho} \in \mathbb{R}^{n\times n}$. It follows from [41] that

$$\nabla E(Z) = 2H(Z)Z. \tag{2.28}$$

The next lemma provides the derivative of the entropy term $R(Z^*Z)$ with respect to Z.

LEMMA 2.3. *Suppose that $Z \in \mathbb{C}^{n\times p}$ satisfies $0 < \sigma_i(Z) < 1$ for all $i = 1,\dots,p$. Let $Z = U\mathrm{Diag}(\sigma)V^*$ for $\sigma = \sigma(Z)$ and some $(U, V) \in \Omega(Z)$. Then the derivative of $R(Z^*Z)$ at Z is*

$$\nabla R(Z^*Z) = 2U\mathrm{Diag}\left(\sigma_1 s'(\sigma_1^2), \dots, \sigma_p s'(\sigma_p^2)\right) V^*, \tag{2.29}$$

where $s'(t) = \ln \frac{t}{1-t}$.

Proof. Let $F = Z^*Z$. It follows from the singular value decomposition of Z that $F = V\mathrm{Diag}\left(\sigma_1^2, \ldots, \sigma_p^2\right) V^*$. Applying the derivatives of spectral functions in [24], we obtain the derivative of $R(F)$ with respect to F:

$$\nabla_F R(F) = \nabla_F \left(\sum_i s(\sigma_i^2) \right) = V\mathrm{Diag}\left(s'(\sigma_1^2), \ldots, s'(\sigma_p^2)\right) V^*,$$

which yields

$$\nabla_Z R(Z^*Z) = 2ZV\mathrm{Diag}\left(s'(\sigma_1^2), \ldots, s'(\sigma_p^2)\right) V^*.$$

This completes the proof. □

3. A Nonmonotone Proximal Gradient Method for EDFT.

In this section, we propose a nonmonotone proximal gradient (NPG) for solving (2.27) inspired by the recent progress on sparse and low-rank matrix optimization. It is an extension of the fixed-point algorithm [18, 43, 44] for minimizing a summation of a Lipschitz continuously differentiable function and a possibly nonsmooth function and the iterative reweighted singular value minimization methods for ℓ_p regularized unconstrained matrix minimization [28].

At the k-th iteration, the proximal gradient method linearizes the term $E(Z)$ with a proximal term but keeps the entropy term $\alpha R(Z^*Z)$. Specifically, the subproblem is

$$\begin{aligned} \min_{Z \in \mathbb{C}^{n\times p}} \quad & \Psi_k(\tau_k; Z) := \langle \nabla E(Z^k), Z - Z^k \rangle + \frac{1}{2\tau_k}\|Z - Z^k\|_F^2 + \alpha R(Z^*Z) \\ \text{s.t.} \quad & \|Z\|_F = \sqrt{p_e}, \\ & 0 \le \sigma_i(Z) \le 1, \quad i = 1, \ldots, p, \end{aligned} \tag{3.1}$$

where τ_k is the proximal step size. Let $W^k = Z^k - \tau_k \nabla E(Z^k)$. The subproblem (3.1) can be expressed as

$$\begin{aligned} \min_{Z \in \mathbb{C}^{n\times p}} \quad & \frac{1}{2\tau_k}\|Z - W^k\|_F^2 + \alpha R(Z^*Z) \\ \text{s.t.} \quad & \|Z\|_F = \sqrt{p_e}, \\ & 0 \le \sigma_i(Z) \le 1, \quad i = 1, \ldots, p. \end{aligned} \tag{3.2}$$

Then the closed-form solution of (3.1) is computable.

LEMMA 3.1. *Given any $W^k \in \mathbb{C}^{n\times p}$ and $\tau_k, \alpha > 0$, let $U\,\mathrm{Diag}(y)V^*$ be the SVD of W^k. Then $Y^k(\tau_k)$ is an optimal solution of* (3.2) *if and only if*

$$Y^k(\tau_k) = U\mathrm{Diag}(z^k(\tau_k))V^* \tag{3.3}$$

and

$$\begin{aligned} z^k(\tau_k) \in \arg\min_z \quad & \psi_k(\tau_k, z) := \frac{1}{2\tau_k}\|z - y\|_F^2 + \alpha \sum_i s(z_i^2) \\ \text{s.t.} \quad & \|z\|_2 = \sqrt{p_e}, \\ & 0 \le z_i \le 1, \quad i = 1, \ldots, p. \end{aligned} \tag{3.4}$$

Proof. The conclusion follows immediately from Lemma 3.1 in [28]. □

Another critical algorithmic issue is the determination of a suitable step size τ. Instead of using the classical Armijo-Wolfe based monotone line search, we apply the nonmonotone curvilinear (as our search path is on the manifold rather than a straight line) search with an initial step size determined by the Barzilai-Borwein (BB) formula, which we have found more efficient for our problem. They were developed originally for the vector case in [3]. At iteration k, the step size is computed as

$$\tau_{k,1} = \frac{\langle S^{k-1}, S^{k-1}\rangle}{|\langle S^{k-1}, V^{k-1}\rangle|} \quad \text{or} \quad \tau_{k,2} = \frac{|\langle S^{k-1}, V^{k-1}\rangle|}{\langle V^{k-1}, V^{k-1}\rangle}, \tag{3.5}$$

where $S^{k-1} = Z^k - Z^{k-1}$ and $V^{k-1} = \nabla E(Z^k) - \nabla E(Z^{k-1})$. In order to guarantee convergence, the final value for τ_k is a fraction (up to 1, inclusive) of $\tau_{k,1}$ or $\tau_{k,2}$ determined by a nonmonotone search condition in [49]. Let $C_0 = \mathcal{M}(Z^0)$, $Q_{k+1} = \eta Q_k + 1$ and $Q_0 = 1$. The new points are generated iteratively in the form $Z_{k+1} := Y^k(\tau_k)$, where $\tau_k = \tau_{k,1}\delta^h$ or $\tau_k = \tau_{k,2}\delta^h$ and h is the smallest nonnegative integer satisfying

$$\mathcal{M}(Y^k(\tau_k)) \le C_k - \frac{\gamma}{2}\|Y^k(\tau_k) - Z^k\|_F^2, \tag{3.6}$$

where $\gamma > 0$, each reference value C_{k+1} is taken to be the convex combination of C_k and $\mathcal{M}(Z^{k+1})$ as $C_{k+1} = (\eta Q_k C_k + \mathcal{M}(Z^{k+1}))/Q_{k+1}$. In Algorithm 2 below, we specify our method for solving the EDFT model. Although several backtracking steps may be needed to update the Z^{k+1}, we observe that the BB step size $\tau_{k,1}$ or $\tau_{k,2}$ is often sufficient for (3.6) to hold in most of our numerical experiments. In the case that $\tau_{k,1}$ or $\tau_{k,2}$ is not bounded, they are reset to a finite number and convergence of our algorithm still holds.

Algorithm 2: A NPG method for EDFT

```
Given $Z^0$, set $\tau_0 > 0$, $\gamma, \delta, \eta \in (0,1)$, $k = 0$, $Q_0 = 1$, $C_0 = \mathcal{M}(Z^0)$.
while convergence is not met do
    for $h = 1, 2, \ldots$ do
        Compute $W^k = X^k - \tau_k \nabla E(Z^k)$ and its SVD.
        Compute $Y^k(\tau_k)$ using (3.3).
        if condition (3.6) is satisfied then break else set the step size $\tau_k = \tau_k \delta^h$
    Set $Z^{k+1} \leftarrow Y^k(\tau_k)$.
    Compute the step size $\tau_{k+1} = \tau_{k+1,1}$ or $\tau_{k+1} = \tau_{k+1,2}$ according to (3.5).
    Update $Q_{k+1} \leftarrow \eta Q_k + 1$ and $C_{k+1} \leftarrow (\eta Q_k C_k + \mathcal{M}(Z^{k+1}))/Q_{k+1}$.
    $k \leftarrow k + 1$.
```

3.1. Convergence of the NPG Method.

The feasible set of (2.27) is denoted by

$$\mathcal{B} = \{Z \mid \|Z\|_F^2 = p_e, \quad 0 \le \sigma_i(Z) \le 1\}.$$

We make the following assumptions.

ASSUMPTION 3.2. *The gradient ∇E is Lipschitz on $\operatorname{conv}\mathcal{B}$, the convex hull of $\mathcal{B}$, with Lipschitz constant $L > 0$.*

Assumption 3.2 holds if there is no exchange-correlation term, for example, in the Gross-Pitaevskii equation [51]. It is also true for the exchange-correlation energy defined by the Perdew & Zunger formula.

LEMMA 3.3. *Let $\tilde{\gamma} = 2\left(\frac{3}{\pi}\right)^{1/3}$ and $(r_s)_i = \left(\frac{4\pi\rho_i}{3}\right)^{-1/3}$. Consider*

$$\epsilon_{xc} = (\epsilon_{ex} + \epsilon_{ec}) \circ \rho, \tag{3.7}$$

where $(\epsilon_{ex})_i = -\frac{3}{4}\tilde{\gamma}\rho_i^{1/3}$ *and*

$$(\epsilon_{ec})_i = \begin{cases} (\epsilon_{ec}^a)_i = A_1 + A_2(r_s)_i + (A_3 + A_4(r_s)_i)\ \ln(r_s)_i & (r_s)_i < 1, \\ (\epsilon_{ec}^b)_i = \frac{B_1}{1+B_2\sqrt{(r_s)_i}+B_3(r_s)_i} & (r_s)_i \geq 1, \end{cases} \tag{3.8}$$

where (with 4 digits accuracy) $A_1 = -0.096, A_2 = -0.0232, A_3 = 0.0622, A_4 = 0.004, B_1 = -0.2846, B_2 = 1.0529$*, and* $B_3 = 0.3334$*. Then the first-order derivative of* ϵ_{xc} *with respect to* Z *is Lipschitz on* $\operatorname{conv}\mathcal{B}$*.*

Proof. Let $(z_i)^\top$ be the i-th row of Z. The definition of ρ gives $\rho_i = \|z_i\|_2^2$. For any $Z \in \operatorname{conv}\mathcal{B}$, we have $1 \geq \sigma_i(Z) = \max_{\|v\|_2=1}\|Zv\|_2 \geq \max_i \|z_i\|_2$, which implies that $\operatorname{conv}\mathcal{B} \subset \{Z\,;\,\rho_i = \|z_i\|^2 \leq 1,\ 1 \leq i \leq n\}$. We first consider the exchange energy term. For $\rho_i \geq 0$, it can be verified that

$$\frac{\partial[(\epsilon_{ex})_i\rho_i]}{\partial\rho_i} = -\tilde{\gamma}\rho_i^{1/3}, \quad \frac{\partial^2[(\epsilon_{ex})_i\rho_i]}{\partial\rho_i^2} = -\frac{1}{3}\tilde{\gamma}\rho_i^{-2/3} \quad (\rho_i > 0).$$

Hence, for $\phi_i(z_i) = (\epsilon_{ex})_i\rho_i\big|_{\rho_i=\|z_i\|^2}$, there holds for all $v, w \in \mathbb{C}^p$:

$$\begin{aligned} &\partial\phi_i(z_i)[v] = -\tilde{\gamma}\|z_i\|^{2/3}(v^*z_i + z_i^*v), \qquad \text{for } z_i \in \mathbb{C}^p, \\ &\partial^2\phi_i(z_i)[v,w] = -\frac{1}{3}\tilde{\gamma}\|z_i\|^{-4/3}(v^*z_i + z_i^*v)(w^*z_i + z_i^*w) - \tilde{\gamma}\|z_i\|^{2/3}(v^*w + w^*v), \qquad \text{for } z_i \in \mathbb{C}^p \setminus \{0\}. \end{aligned}$$

Consequently, it holds $\phi_i(0) = 0$, $\partial\phi_i(0) = 0$, and $\partial^2\phi_i(z_i) \to 0$ as $z_i \to 0$. For $0 \neq h \to 0$, we obtain

$$\|\partial\phi_i(h)[v] - \partial\phi_i(0)[v]\| = \tilde{\gamma}\|h\|^{2/3}\|v^*h + h^*v\| = \|v\|\,O(\|h\|^{5/3}) = \|v\|\,o(\|h\|),$$

which yields $\partial^2\phi_i(0) = 0$. Hence, $\partial^2\phi_i$ is continuous on $\mathbb{C}^p$ and thus it is bounded on $\{z_i \in \mathbb{C}^p\,;\,\|z_i\|^2 \leq 1\}$.

We now consider the correlation energy term. Since $\rho_i \leq 1$, we have $(r_s)_i \geq \left(\frac{4\pi}{3}\right)^{-1/3} =: c > 0$. The two branches $(\epsilon_{ec}^a)_i$ and $(\epsilon_{ec}^b)_i$ of $(\epsilon_{ec})_i$, see (3.8), are smooth functions of $(r_s)_i$ in neighborhoods of $[c, 1]$ and $[1, \infty)$, respectively. All derivatives of $(\epsilon_{ec}^a)_i$ are bounded on $[c, 1]$ and, for any fixed $C > 1$, all derivatives of $(\epsilon_{ec}^b)_i$ are bounded on $[1, C]$. Hence, $\frac{\partial[(\epsilon_{ec})_i]}{\partial(r_s)_i}$ is Lipschitz on $[c, C]$ for any finite $C > 1$. Therefore, to show that the gradient of $\psi_i(z_i) := (\epsilon_{ec})_i\rho_i\big|_{\rho_i=\|z_i\|^2}$ is continuously differentiable on a neigborhood of $\{z_i\,;\,\|z_i\|^2 \leq 1\}$, it is sufficient to consider a neighborhood of $z_i = 0$ as $(r_s)_i$ is unbounded only for $z_i \to 0$. For $\delta > 0$ sufficiently small there holds $1 < (r_s)_i < \infty$ for all $0 < \|z_i\| < \delta$. A short calculation shows that

$$\begin{aligned} \frac{\partial[(\epsilon_{ec})_i\rho_i]}{\partial\rho_i} &= \frac{B_1\,(1 + \frac{7}{6}B_2\sqrt{(r_s)_i} + \frac{4}{3}B_3(r_s)_i)}{(1 + B_2\sqrt{(r_s)_i} + B_3(r_s)_i)^2}, \\ \frac{\partial^2[(\epsilon_{ec})_i\rho_i]}{\partial\rho_i^2} &= \frac{B_1\pi\,(5B_2 + (7B_2^2 + 8B_3)\sqrt{(r_s)_i} + 21B_2B_3(r_s)_i + 16B_3^2(r_s)_i^{3/2})(r_s)_i^{7/2}}{27(1 + B_2\sqrt{(r_s)_i} + B_3(r_s)_i)^3}. \end{aligned}$$

For $0 \neq z_i \to 0$, we obtain $\rho_i \to 0$, $(r_s)_i \to \infty$, and

$$\left|\frac{\partial[(\epsilon_{ec})_i\rho_i]}{\partial\rho_i}\right| = O((r_s)_i^{-1}) = O(\rho_i^{1/3}), \quad \left|\frac{\partial^2[(\epsilon_{ec})_i\rho_i]}{\partial\rho_i^2}\right| = O((r_s)_i^2) = O(\rho_i^{-2/3}).$$

Thus, for $z_i \to 0$ and all $v, w \in \mathbb{C}^p$, we have

$$\begin{aligned}
&|\partial\psi_i(z_i)[v]| = \left|\frac{\partial[(\epsilon_{ec})_i\rho_i]}{\partial\rho_i}(z_i^* v + v^* z_i)\right| = \|v\|\, O(\rho_i^{1/3}\|z_i\|) = \|v\|\, O(\|z_i\|^{5/3}) = \|v\|\, o(\|z_i\|),\\
&|\partial^2\psi_i(z_i)[v,w]| = \left|\frac{\partial[(\epsilon_{ec})_i\rho_i]}{\partial\rho_i}(w^* v + v^* w) + \frac{\partial^2[(\epsilon_{ec})_i\rho_i]}{\partial\rho_i^2}(z_i^* v + v^* z_i)(z_i^* w + w^* z_i)\right|\\
&= \|v\|\|w\|\, O(\rho_i^{1/3} + \rho_i^{-2/3}\|z_i\|^2) = \|v\|\|w\|\, O(\|z_i\|^{2/3}) \to 0,
\end{aligned}$$

which implies that $\partial^2\psi_i$ is bounded in a neighborhood of 0 and $\partial^2\psi_i(0) = 0$. Hence, $\partial\psi_i$ is Lipschitz on $\{z_i\,;\, \|z_i\| \leq 1\}$. □

We next show that $z^k(\tau_k)$ of (3.4) does not lie on the boundary of $[0,1]$. Note that $y \geq 0$ since the components of y are singular values.

LEMMA 3.4. *For any $y \in \mathbb{R}_+^p$ and $\tau_k > 0$, the optimal solution $z^k(\tau_k)$ of* (3.4) *satisfies $0 < z_i^k(\tau_k) < 1$ for $i = 1, \ldots, p$.*

Proof. For brevity, $z^k(\tau_k)$ is denoted by $\hat{z}$.

1) We first prove $\hat{z}_i > 0$ for all i. Assume that there exists an index i such that $\hat{z}_i = 0$. Since $p_e > 0$, the constraints of (3.4) imply that there exists an index j such that $0 < \hat{z}_j \leq 1$. Define a vector $w(t)$ such that

$$w_l(t) = \begin{cases} \hat{z}_l, & l \neq i, j,\\ t, & l = i,\\ \sqrt{\hat{z}_j^2 - t^2}, & l = j, \end{cases}$$

where $t > 0$ is sufficiently small. It holds that $w_i(t)^2 + w_j(t)^2 = \hat{z}_i^2 + \hat{z}_j^2 = \hat{z}_j^2$. Hence, $w(t)$ is feasible for all $t \in [0,\delta]$ with $\delta > 0$ sufficiently small and there holds $w_i(t), w_j(t) \in (0,1)$ for all $t \in (0,\delta)$. Further, $\phi(t) := \psi_k(\tau_k; w(t))$ is smooth on $(0,\delta)$ and continuous on $[0,\delta]$. For $t \in (0,\delta)$, we obtain

$$\phi'(t) = \frac{1}{\tau_k}(t - y_i) + \frac{1}{\tau_k} w_j'(t)(w_j(t) - y_j) + \alpha s'(t^2)\cdot 2t + \alpha s'(w_j(t)^2)\cdot 2w_j'(t)w_j(t).$$

There holds

$$w_j'(t) = -t/\sqrt{\hat{z}_j^2 - t^2}, \quad w_j'(t) = -t/\hat{z}_j + o(t), \text{ as } t \to 0^+.$$

It can be verified that

$$\begin{aligned}
&\lim_{t\to 0^+} \frac{1}{t}\,\frac{1}{\tau_k}(t - y_i) = \begin{cases} 1/\tau_k, & y_i = 0,\\ -\infty, & y_i > 0,\end{cases}\\
&\lim_{t\to 0^+} \frac{1}{t}\,\frac{1}{\tau_k} w_j'(t)(w_j(t) - y_j) = -\frac{1}{\tau_k}\frac{\hat{z}_j - y_j}{\hat{z}_j},\\
&\lim_{t\to 0^+} \frac{1}{t}\,\alpha s'(t^2)\cdot 2t = -\infty,\\
&\lim_{t\to 0^+} \frac{1}{t}\,\alpha s'(w_j(t)^2)\cdot 2w_j'(t)w_j(t) = \begin{cases} -2\alpha s'(\hat{z}_j^2) \in \mathbb{R}, & \hat{z}_j < 1,\\ -\infty, & \hat{z}_j = 1.\end{cases}
\end{aligned}$$

The previous relationships show $\phi'(t)/t \to -\infty$. Hence, there exists $0 < \delta_1 < \delta$ such that $\phi'(t) < 0$ for all $t \in (0,\delta_1]$.

Then $\psi_k(\tau_k; z(t)) = \phi(t) < \phi(0) = \psi_k(\tau_k; \hat{z})$, which shows that $\hat{z}$ is not optimal. This is a contradiction.

2) We now prove $\hat{z}_i < 1$ for all i. Assume that there exists an index i such that $\hat{z}_i = 1$. Since $p_e < p$, the constraints of (3.4) imply that there exists another index j with $0 \le \hat{z}_j < 1$. By part 1) we know in addition that $\hat{z}_j > 0$. Define a vector $w(t)$ as

$$w_l(t) = \begin{cases} \hat{z}_l, & l \neq i, j, \\ 1 - t, & l = i, \\ \sqrt{\hat{z}_j^2 + 2t - t^2}, & l = j, \end{cases}$$

where $t > 0$ is sufficiently small. It is obvious that $w_i^2(t) + w_j^2(t) = \hat{z}_i^2 + \hat{z}_j^2 = 1 + \hat{z}_j^2$. Hence, $w(t)$ is feasible for all $t \in [0, \delta]$ if $\delta > 0$ is sufficiently small and there holds $w_i(t), w_j(t) \in (0, 1)$ for all $t \in (0, \delta)$. Further, $\phi(t) := \psi_k(\tau_k, w(t))$ is smooth on $(0, \delta)$ and continuous on $[0, \delta]$. For $t \in (0, \delta)$, we obtain

$$\phi'(t) = -\frac{1}{\tau_k}(1 - t - y_i) + \frac{1}{\tau_k} w_j'(t)(w_j(t) - y_j) + \alpha s'((1-t)^2) \cdot 2(t-1) + \alpha s'(w_j(t)^2) \cdot 2w_j'(t) w_j(t).$$

Since $\hat{z}_j > 0$, we have $w_j'(t) \to w_j'(0) = 1/\hat{z}_j$ as $t \to 0^+$ and

$$\begin{aligned} & -\frac{1}{\tau_k}(1 - t - y_i) + \frac{1}{\tau_k} w_j'(t)(w_j(t) - y_j) + \alpha s'(w_j(t)^2) \cdot 2w_j'(t) w_j(t) \\ & \to -\frac{1}{\tau_k}(1 - y_i) + \frac{1}{\tau_k}\frac{1}{\hat{z}_j}(\hat{z}_j - y_j) + 2\alpha s'(\hat{z}_j^2), \end{aligned}$$

whose value is finite. Furthermore, it holds $\alpha s'((1-t)^2) \cdot 2(t-1) \to -\infty$, which shows $\phi'(t) \to -\infty$ as $t \to 0^+$.

Therefore, due to the smoothness of ϕ on $(0, \delta)$, there exists $0 < \delta_1 < \delta$ such that $\phi'(t) < 0$ for all $t \in (0, \delta_1]$ and $\psi_k(\tau_k; w(t)) = \phi(t) < \phi(0) = \psi_k(\tau_k; \hat{z})$, which shows that $\hat{z} = z_k(\tau_k)$ is not optimal. This is a contradiction. □

LEMMA 3.5. *Given two sequences $\{y^k\} \subset \mathbb{R}_+^p$ and $\{\tau^k\} \subset \mathbb{R}_{++}$. Let $\{\hat{z}^k\} = z^k(\tau_k)$ be the corresponding solutions of* (3.4) *by replacing y by y_k. Consider any subsequence $\mathcal{K}$ such that $\{y^k\}_{\mathcal{K}} \to \bar{y}$, $\{\tau^k\}_{\mathcal{K}} \to \bar{\tau} > 0$, and $\{\hat{z}^k\}_{\mathcal{K}} \to \bar{z}$. Then there holds $\bar{y} \in \mathbb{R}_+^p$ and $\bar{z}$ solves* $(3.4)_{y \leftarrow \bar{y}, \tau_k \leftarrow \bar{\tau}}$ *(i.e., y and τ_k are replaced by $\bar{y}$ and $\bar{\tau}$, respectively). In particular, there holds $0 < \bar{z} < 1$.*

Proof. Let us define $\mathcal{F} = \{z \mid \|z\|_2^2 = p_e, \quad 0 \le z \le 1\}$ and $g : \mathbb{R}^p \to \mathbb{R}^p \cup \{+\infty\}$,

$$g(z) := \begin{cases} \alpha \sum_i s(z_i^2), & \|z\|_2^2 = p_e, \ 0 \le z \le 1, \\ +\infty, & \text{otherwise.} \end{cases}$$

Then the domain $\text{dom}\,(g) = \{z \mid g(z) < \infty\} = \mathcal{F}$ is nonempty and compact and g is continuous on $\text{dom}\,(g)$. Therefore, g is proper, lower semicontinuous and proximal-bounded (for all thresholds $\tau > 0$, see Def 1.23 and Ex. 1.24 in [35]). The proximal operator of g using the parameter $\tau > 0$ is defined as the set-valued map

$$\text{prox}_\tau^g(y) = \underset{z \in \mathbb{R}^p}{\text{argmin}}\ g(z) + \frac{1}{2\tau}\|z - y\|_2^2.$$

According to Theorem 1.25 in [35], it holds that for all $\tau > 0$ and all $y \in \mathbb{R}^p$ the set $\text{prox}_\tau^g(y)$ is nonempty and compact. Further, if $0 < \tau_\nu \to \tau > 0$, $y_\nu \to y$, and $w_\nu \in \text{prox}_{\tau_\nu}^g(y_\nu)$, then all cluster points of $\{w_\nu\}$ are contained in $\text{prox}_\tau^g(y)$.

It can be easily seen that $\hat{z}^k$ solves $(3.4)_{y \leftarrow y^k}$ if and only if $\hat{z}^k \in \text{prox}_{\tau_k}^g(y^k)$. By the above continuity properties of the proximal operator, we conclude from $\{y^k\}_{\mathcal{K}} \to \bar{y}$, $\{\hat{z}^k\}_{\mathcal{K}} \to \bar{z}$, $\{\tau_k\}_{\mathcal{K}} \to \bar{\tau} > 0$ that $\bar{z} \in \mathrm{prox}^g_{\bar{\tau}}(\bar{y})$. Hence, $\bar{z}$ solves $(3.4)_{y \leftarrow \bar{y}, \tau_k \leftarrow \bar{\tau}}$ and $y_k \geq 0$ implies $\bar{y} \geq 0$. It follows from Lemma 3.4 that $0 < \bar{z} < 1$. □

The next lemma proves that the Armijo condition (3.6) holds after a finite number of back-tracking searches.

LEMMA 3.6. *Suppose that Assumption 3.2 holds. For the sequence* $\{Z^k\}$ *generated by Algorithm 2, we have* $\mathcal{M}(Z^k) \leq C_k \leq a_k$ *and* $Q_k \leq 1/(1-\eta)$, *where* $a_k = \frac{1}{k+1}\sum_{i=0}^{k} \mathcal{M}(Z^k)$. *The curvilinear condition* (3.6) *is satisfied if* $\tau_k \in (0, 1/(L+\gamma)]$.

Proof. The inequalities $\mathcal{M}(Z^k) \leq C_k \leq a_k$ and $Q_k \leq 1/(1-\eta)$ follow directly from the proof of Lemma 1.1 and Theorem 2.2 in [49]. For all $Z \in \mathrm{conv}(\mathcal{B})$ and $D = Z - Z^k$, we have

$$\begin{aligned} E(Z^k + D) - E(Z^k) - \langle \nabla E(Z^k), D \rangle &= \int_0^1 \langle \nabla E(Z^k + tD) - \nabla E(Z^k), D \rangle \, dt \\ &\leq \|D\|_F \int_0^1 Lt \|D\|_F \, dt = \frac{L}{2}\|D\|_F^2. \end{aligned}$$

Using $Y^k(\tau_k)$ is an optimal solution of (3.1) and $Z^k \in \mathcal{B}$ yields

$$\begin{aligned} \alpha R((Z^k)^* Z^k) &= \Psi_k(\tau_k; Z^k) \geq \Psi_k(\tau_k; Y^k(\tau_k)) \\ &= \langle \nabla E(Z^k), Y^k(\tau_k) - Z^k \rangle + \frac{1}{2\tau_k}\|Y^k(\tau_k) - Z^k\|_F^2 + \alpha R(Y^k(\tau_k)^* Y^k(\tau_k)) \\ &\geq E(Y^k(\tau_k)) - E(Z_k) - \frac{L}{2}\|Y^k(\tau_k) - Z_k\|_F^2 + \frac{1}{2\tau_k}\|Y^k(\tau_k) - Z_k\|_F^2 + \alpha R(Y^k(\tau_k)^* Y^k(\tau_k)). \end{aligned}$$

Rearranging terms yields

$$\mathcal{M}(Y^k(\tau_k)) - \mathcal{M}(Z^k) \leq \left(\frac{L}{2} - \frac{1}{2\tau_k}\right)\|Y^k(\tau_k) - Z^k\|_F^2. \tag{3.9}$$

Hence, (3.6) holds since $\tau_k \in (0, 1/(L+\gamma)]$ and $\mathcal{M}(Z^k) \leq C_k$. □

The next lemma gives the first-order optimality conditions of (2.27).

LEMMA 3.7. *For any local minimizer* $\bar{Z} \in \mathbb{C}^{n \times p}$ *of* (2.27), *there exists a Lagrangian multiplier* λ *such that the following first-order optimality conditions hold:*

$$H(\bar{Z})\bar{Z} + U\mathrm{Diag}\left(\pi(\sigma(\bar{Z}))\right)V^* = 0, \ \textit{for some}(U, V) \in \Omega(\bar{Z}). \tag{3.10}$$

$$\|\bar{Z}\|_F = \sqrt{p_e}, \tag{3.11}$$

$$0 < \sigma(\bar{Z}) < 1, \tag{3.12}$$

where $\pi_i(\sigma_i) = 2\alpha\sigma_i s'(\sigma_i^2) - \lambda\sigma_i$, $i = 1, \ldots, p$.

Proof. We first show that $\bar{Z}$ is a global minimizer of the problem

$$\begin{aligned} \min_{Z \in \mathbb{C}^{n \times p}} \quad & \frac{1}{2\bar{\tau}}\|Z - \bar{W}\|_F^2 + \alpha R(Z^* Z) \\ \text{s.t.} \quad & \|Z\|_F = \sqrt{p_e}, \\ & 0 \leq \sigma_i(Z) \leq 1, \quad i = 1, \ldots, p, \end{aligned} \tag{3.13}$$

where $\bar{W} = \bar{Z} - \bar{\tau}\nabla E(\bar{Z})$, and $\bar{\tau} \in (0, 1/(L+\gamma)]$. Let $\bar{Z}(\bar{\tau})$ be a global solution of (3.13). The proof of Lemma 3.6

shows that

$$\mathcal{M}(\bar{Z}(\bar{\tau})) \leq \mathcal{M}(\bar{Z}) - \frac{\gamma}{2}\|\bar{Z}(\bar{\tau}) - \bar{Z}\|_F^2.$$

which implies that $\mathcal{M}(\bar{Z}(\bar{\tau})) < \mathcal{M}(\bar{Z})$ if $\bar{Z}(\bar{\tau}) \neq \bar{Z}$. Since $R(Z^*Z)$ is bounded below on the feasible set of (3.13), there holds $\bar{Z}(\bar{\tau}) \to \bar{Z}$ as $\bar{\tau} \to 0^+$. Thus, for all sufficiently small $\bar{\tau} > 0$ we must have $\bar{Z}(\bar{\tau}) = \bar{Z}$ since otherwise $\mathcal{M}(\bar{Z}(\bar{\tau})) < \mathcal{M}(\bar{Z})$ would contradict the fact that $\bar{Z}$ is a local minimizer of (2.27).

Applying Lemma 3.4 to (3.13) gives $0 < \sigma_i(\bar{Z}) < 1$. Hence, $\bar{Z}$ is also a local minimizer of

$$\begin{aligned} \min_{Z \in \mathbb{C}^{n\times p}} \quad & E(Z) + \alpha R(Z^*Z) \\ \text{s.t.} \quad & \|Z\|_F = \sqrt{p_e}. \end{aligned} \tag{3.14}$$

Since the linearly independent constraint qualification (LICQ) holds at $\bar{Z}$, the first-order optimality conditions of (3.14) yields (3.10)-(3.12). □

We next establish the convergence results of Algorithm 2.

THEOREM 3.8. *Let $\{Z^k\}$ be a sequence generated by Algorithm 2. Then the sequence $\{Z^k\}$ is bounded and any accumulation point $\bar{Z}$ of $\{Z^k\}$ satisfies the first-order optimality conditions* (3.10)-(3.12) *of problem* (2.27).

Proof. The boundedness of $\{Z^k\}$ follows from the fact that $Z^k \in \mathcal{B}$. From the updating rule C_{k+1} and (3.6) , we obtain

$$C_{k+1} = \frac{\eta Q_k C_k + \mathcal{M}(Z^{k+1})}{Q_{k+1}} \leq \frac{\eta Q_k C_k + C_k - \frac{\gamma}{2}\|Z^{k+1} - Z^k\|_F^2}{Q_{k+1}} = C_k - \frac{\gamma\|Z^{k+1} - Z^k\|_F^2}{2Q_{k+1}}.$$

Since $\{Z^k\}$ is bounded and $\mathcal{M}(Z^k) \leq C_k$ for all k, the function value $\mathcal{M}(Z^k)$ and C_k are bounded from below. Hence, we obtain $\sum_{k=0}^{\infty} \frac{\|Z^{k+1}-Z^k\|_F^2}{Q_{k+1}} < \infty$, which together with the fact $Q_{k+1} \leq 1/(1-\eta)$ from Lemma 3.6 implies that $\lim_{k\to\infty}\|Z^{k+1} - Z^k\|_F = 0$.

Let $\bar{\tau}_k$ denote the final value of τ_k such that (3.6) holds. It follows from Lemma 3.6 that $\{\bar{\tau}_k\}$ is bounded away from 0 by $\delta/(L+\gamma)$. As observed from Algorithm 2, the point Z^{k+1} is the solution of the subproblem (3.1) with $\tau_k = \bar{\tau}_k$, i.e.,

$$\begin{aligned} Z^{k+1} = \arg\min_{z} \quad & \frac{1}{2\bar{\tau}_k}\|Z - W^k\|_F^2 + \alpha R(Z^*Z) \\ \text{s.t.} \quad & \|Z\|_F = \sqrt{p_e}, \\ & 0 \leq \sigma_i(Z) \leq 1, \quad i = 1, \ldots, p, \end{aligned} \tag{3.15}$$

where $W^k = Z^k - \bar{\tau}_k \nabla E(Z^k)$. Let $U^k \mathrm{Diag}(y^k)(V^k)^*$ be the SVD of W^k. By applying Lemma 3.1, we obtain

$$Z^{k+1} = U^k \mathrm{Diag}(z^{k+1})(V^k)^*, \tag{3.16}$$

where

$$\begin{aligned} z^{k+1} = \arg\min_{z} \quad & \frac{1}{2\bar{\tau}_k}\|z - y^k\|_F^2 + \alpha \sum_i s(z_i^2) \\ \text{s.t.} \quad & \|z\|_2 = \sqrt{p_e}, \\ & 0 \leq z_i \leq 1, \quad i = 1, \ldots, p. \end{aligned} \tag{3.17}$$

Using Lemma 3.4, we obtain the first-oder optimality conditions of (3.15) as

$$\begin{aligned} \frac{1}{\bar{\tau}_k}(Z^{k+1}-Z^k)+\nabla E(Z^k)+U^k\mathrm{Diag}\left(\pi^k(z^{k+1})\right)(V^k)^* = 0, \\ \|Z^k\|_F = \sqrt{p_e}, \\ 0 < z_i^{k+1} < 1, \quad i=1,\dots,p, \end{aligned} \tag{3.18}$$

where $\pi_i^k(z_i^{k+1}) = 2\alpha z_i^{k+1} s'((z_i^{k+1})^2) - \lambda_k z_i^{k+1}$, $i=1,\dots,p$.

Let $\bar{Z}$ be an accumulation point of $\{Z^k\}$. Then there exists a subsequence $\mathcal{K}$ such that $\{Z^k\}_{\mathcal{K}} \to \bar{Z}$ and $\{\bar{\tau}^k\}_{\mathcal{K}} \to \bar{\tau} > 0$, which together with $\lim_{k\to\infty}\|Z^{k+1}-Z^k\|_F = 0$ implies that $\{Z^{k+1}\}_{\mathcal{K}} \to \bar{Z}$. Since $Z \mapsto \sigma(Z)$ is continuous, we have $\{z^{k+1}\}_{\mathcal{K}} = \{\sigma(Z^{k+1})\}_{\mathcal{K}} \to \sigma(\bar{Z}) =: \bar{z}$. From Lemma 3.5 we obtain $0 < \bar{z} < 1$. The first equation in (3.18) shows that $\{\lambda_k z^{k+1}\}$ is bounded since all other terms in this equation are bounded and U^k, V^k are unitary. From $\{z^{k+1}\}_{\mathcal{K}} \to \bar{z} > 0$ we obtain that $\{\lambda_k\}_{\mathcal{K}}$ is bounded. Hence, we can select a subsequence $\mathcal{K}' \subset \mathcal{K}$ such that $\{\lambda_k\}_{\mathcal{K}'}$ converges to some limit $\lambda \in \mathbb{R}$. Define $\pi(z)$ by $\pi_i(z_i) = 2\alpha z_i s'((z_i)^2) - \lambda z_i$, $i=1,\dots,p$. It follows from $0 < \bar{z} < 1$ that the function π is continuous at $\sigma(\bar{Z}) = \bar{z}$. Thus, the spectral operator $Z = U\mathrm{Diag}(\sigma(Z))V^* \mapsto U\mathrm{Diag}\left(\pi(\sigma(Z))\right)V^*$ is continuous at $\bar{Z}$ by Theorem 3.1 in [13]. Taking the limit $\mathcal{K}' \ni k \to \infty$ yields:

$$\begin{aligned} &U^k\mathrm{Diag}\left(\pi^k(z^{k+1})\right)(V^k)^* = U^k\mathrm{Diag}\left(\pi(\sigma(Z^{k+1}))\right)(V^k)^* + U^k\mathrm{Diag}\left(\pi^k(\sigma(Z^{k+1})) - \pi(\sigma(Z^{k+1}))\right)(V^k)^* \\ &\to \bar{U}\mathrm{Diag}\left(\pi(\sigma(\bar{Z}))\right)\bar{V}^*, \end{aligned}$$

where we have used the SVD $Z^{k+1} = U^k\mathrm{Diag}(z^{k+1})(V^k)^*$ with $\sigma(Z^{k+1}) = z^{k+1}$, the boundedness of $\{U^k\}$ and $\{V^k\}$, the continuity of π at $\sigma(\bar{Z})$, and $|\pi^k(z^{k+1}) - \pi(z^{k+1})| = \|(\lambda_k - \lambda)z^{k+1}\| \to 0$ as $\mathcal{K}' \ni k \to \infty$. Hence, the optimality conditions (3.10)-(3.12) are derived by taking limits $\mathcal{K}' \ni k \to \infty$ in (3.18). □

3.2. Solving the NPG Subproblem (3.4)**.** Consider the optimization problem

$$\begin{aligned} \min_{x\in\mathbb{R}^p} \quad & \frac{1}{2\tau}\|x-u\|^2 + \sum_{i=1}^p \left(x_i^2 \ln x_i^2 + (1-x_i^2)\ln(1-x_i^2)\right) \\ \text{s.t.} \quad & \|x\|^2 = r, \quad 0 \le x \le 1, \end{aligned} \tag{3.19}$$

where $\tau > 0$ and $u \ge 0$. For example, (3.4) is a special case of (3.19) by choosing $\tau = \tau_k\alpha$, $u = y$ and $r = p_e$. Using Lemma 3.4, it is easy to verify that the linearly independent constraint qualification (LICQ) holds at any local solution of (3.19). The optimal solution of (3.19) can be obtained by finding λ^* such that

$$\|x^*(\lambda^*)\|^2 - r = 0, \tag{3.20}$$

where $x^*(\lambda)$ is the optimal solution of

$$\begin{aligned} \min_{x} \quad & \frac{1}{2\tau}\|x-u\|^2 + \sum_{i=1}^p \left(x_i^2 \ln x_i^2 + (1-x_i^2)\ln(1-x_i^2)\right) + \frac{\lambda}{2}\|x\|^2 \\ \text{s.t.} \quad & 0 \le x \le 1. \end{aligned} \tag{3.21}$$

We next show that $x^*(\lambda)$ is unique for a given λ and its derivative with respect to λ can be computed explicitly. Hence, Newton's method can be used to find the roots λ^* for (3.20).

Since problem (3.21) is separable in each variable x_i, it follows that

$$x_i^*(\lambda) = \arg\min_{0\le x_i\le 1} \left(\frac{1}{2\tau}+\frac{\lambda}{2}\right)x_i^2 - \frac{1}{\tau}u_i x_i + x_i^2\ln x_i^2 + (1-x_i^2)\ln(1-x_i^2), \quad i=1,\ldots,p. \tag{3.22}$$

It is suffice to consider the problem

$$\min_{0\le y\le 1} ay^2 + by + y^2\ln y^2 + (1-y^2)\ln(1-y^2), \tag{3.23}$$

where $b \le 0$. Upon letting $t = y^2$, problem (3.23) is equivalent to

$$t^* = \arg\min_{0\le t\le 1} \underbrace{at + b\sqrt{t} + t\ln t + (1-t)\ln(1-t)}_{\phi(t)}. \tag{3.24}$$

We next discuss how to find such a t^* in (3.24). Observe that

$$\begin{aligned}\phi'(t) &= a + \frac{bt^{-1/2}}{2} + \ln t - \ln(1-t),\\ \phi''(t) &= \frac{t^{-1/2}}{4t(1-t)}[4t^{1/2} - b(1-t)].\end{aligned} \tag{3.25}$$

Define

$$\phi'_+(t)(0) = \lim_{t\to 0^+}\phi'(t), \quad \phi'_-(t)(1) = \lim_{t\to 1^-}\phi'(t).$$

Since $b \le 0$, one can observe that

$$\phi'_+(t)(0) = -\infty, \quad \phi'_-(t)(1) = \infty, \quad \phi''(t) > 0 \ \ \forall t \in (0,1).$$

It follows that $t^* \ne 0$ or 1 and t^* is the unique root of the equation $\phi'(t) = 0$ in $(0,1)$, which can be found by applying the Newton's method or bisection method to this equation. This completes the proof that $x^*(\lambda)$ is unique for a given λ.

We next show that $x^*(\lambda)$ is differentiable with respect to λ by studying the relationship between the parameter a and the solution t^* of problem (3.24). To emphasize the dependence of t^* on a, we denote t^* and $\phi(t)$ by $t^*(a)$ and $\tilde{\Phi}(a,t)$, respectively. Namely,

$$\begin{aligned}\Phi(a,t) &:= at + b\sqrt{t} + t\ln t + (1-t)\ln(1-t)\\ t^*(a) &:= \arg\min_{0\le t\le 1}\Phi(a,t).\end{aligned}$$

Note that $\Phi(a,0) = 0$ and $\Phi(a,t^*(a)) = \phi(t^*(a))$. Moreover, $\Phi(a,t^*(a)) < 0$ since $t^*(a) > 0$ and $\Phi(a,\cdot)$ is strictly decreasing for t close to zero.

PROPOSITION 3.9. *There hold*

(i) $t^*(a)$ *is strictly decreasing and differentiable in* $(-\infty,+\infty)$.
(ii) $\lim_{a\to-\infty} t^*(a) = 1$ *and* $\lim_{a\to+\infty} t^*(a) = 0$.

Proof. (i) Since $b \le 0$, $t^*(a)$ is the unique root of $\phi'(t) = 0$ in $(0,1)$ for any $a \in \mathbb{R}$. In other words, $(a, t^*(a))$

satisfies

$$\tilde{\Phi}(a,t) = \frac{\partial \Phi}{\partial t} = a + \frac{bt^{-1/2}}{2} + \ln t - \ln(1-t) = 0. \tag{3.26}$$

Observe that $\tilde{\Phi}$ is differentiable in $\mathbb{R} \times (0,1)$ and moreover

$$\frac{\partial \tilde{\Phi}}{\partial t} = \frac{t^{-1/2}}{4t(1-t)}[4t^{1/2} - b(1-t)] > 0$$

for any $t \in (0,1)$ due to $b \leq 0$. Invoking the implicit function theorem, one can conclude that $t^*(a)$ is differentiable in $\mathbb{R}$. Differentiating both sides of (3.26) with respect to a and letting $t = t^*(a)$, we have

$$1 + \frac{\partial \tilde{\Phi}}{\partial t}\Big|_{(a,t^*(a))} \cdot \frac{dt^*(a)}{da} = 0.$$

This relation and $\frac{\partial \tilde{\Phi}}{\partial t}\Big|_{(a,t^*(a))} > 0$ yield $\frac{dt^*(a)}{da} < 0$. Hence $t^*(a)$ is strictly decreasing in $\mathbb{R}$.

(ii) Since $t^*(a)$ is monotone and $t^*(a) \in (0,1)$, $\lim_{a\to-\infty} t^*(a)$ and $\lim_{a\to\infty} t^*(a)$ must exist and they are some numbers in $[0,1]$. We know that $(a, t^*(a)) \in \mathbb{R} \times (0,1)$ satisfies (3.26). Hence, one has

$$a + \frac{b}{2\sqrt{t^*(a)}} + \ln t^*(a) - \ln(1 - t^*(a)) = 0. \tag{3.27}$$

Using this equality, $\lim_{a\to-\infty} t^*(a) \in [0,1]$ and $\lim_{a\to\infty} t^*(a) \in [0,1]$, we can see that the conclusion holds. □

It follows from Proposition 3.9 that $x^*(\lambda)$ of (3.21) satisfies the following properties.

PROPOSITION 3.10. *There hold*

(i) $\|x^*(\lambda)\|^2$ *is continuous and decreasing in* $(-\infty, \infty)$.

(ii) $\lim\limits_{\lambda\to-\infty} x^*(\lambda) = e$ *and* $\lim\limits_{\lambda\to+\infty} x^*(\lambda) = 0$, *where* e *is the all-ones vector.*

4. Numerical Results. In order to demonstrate the effectiveness of our proximal gradient method, we implemented the method within KSSOLV [46], which is a MATLAB toolbox for solving the Kohn-Sham problem. In KSSOLV, the Kohn-Sham problem is discretized by using planewave expansions. The number of planewaves used to expand each wavefunction, denoted by n_g, is determined by a kinetic energy cutoff E_{cut}. Each wavefunction is uniformly sampled on a $n_1 \times n_2 \times n_3$ grid on which the charge density ρ as well as part of the potential are evaluated. When computing the gradient of $E(Z)$, we apply the kinetic energy operator L and nonlocal potential W in the Fourier space and the local potential in real space. Fast fourier transforms (FFTs) are used to convert from one representation of the wavefunction (Z) to the other.

We compare the efficiency and accuracy of the proximal gradient method with a finite-temperature version of the SCF iteration. All the experiments were performed on a Dell Precision T7600 workstation with Intel Xenon(R) E5-2697 CPU at 2.70GHz (×12) and 128GB of memory running Ubuntu 12.04 and MATLAB 2013b. For each Z^k computed by NPG, we can recover an orthogonal matrix X^k using the second statement of Theorem 2.1. All methods were terminated if the residual $\|HX - X(X^*HX)\|_F$ is less than 10^{-5}. The linear eigenvalue problems in SCF were solved by LOBPCG [20] which itself is an iterative method. We stop LOBPCG when either the maximum residual norm among all desired eigenpairs is below 10^{-6} or a maximal of 10 iterations is reached.

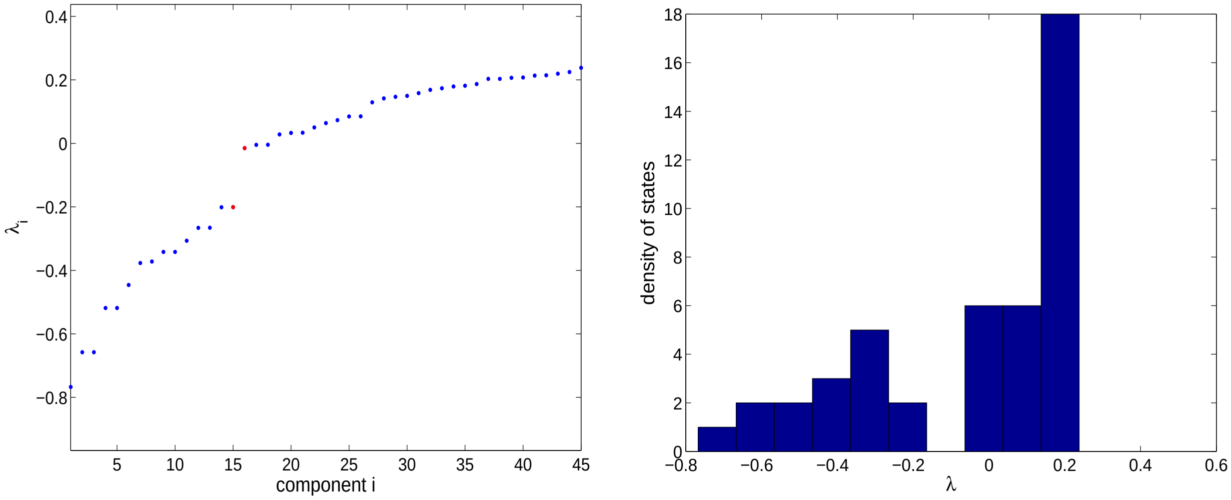
The test problems we choose include both insulating systems (benzene, alanine, hnco, si2h4) and metallic systems (ctube661, graphene16, ptnio, nic). The sizes of the Hamiltonians and wavefunctions associated with with examples are summarized in Table 1. We also list the number of electron pairs p_e. Since we ignore spins in this paper, p_e is the same as the number of occupied states in a zero temperature DFT calculation.

In a finite temperature ensemble DFT calculation, we need to keep some additional wavefunctions since we allow more partially occupied wavefunctions to be included in the definition of the charge density. We typically take the total number of wavefunctions to be included in the charge density (p) to be three times the number of electron pairs. For insulating systems, the occupation number associated with the $p+1$ wavefunction typically has a tiny fraction of occupation at low temperature. Thus one can include fewer extra wavefunctions in the charge density construction.

TABLE 1
Problem Information

name	(n_1, n_2, n_3)	n	p_e
benzene	(64, 64, 32)	8407	15
alanine	(64, 48, 64)	12671	18
hnco	(32, 32, 32)	2103	8
si2h4	(32, 32, 32)	2103	6
ctube661	(115, 115, 15)	12599	48
graphene16	(57, 57, 15)	3071	37
ptnio	(63, 34, 30)	4069	43
nic	(16, 16, 16)	251	7

An insulating system typically has a relatively large gap between the p_eth eigenvalue and p_e+1st eigenvalue. This can be seen from the distribution of the first 45 eigenvalues of the benzene Hamiltonian at T = 1000K in Figure 1. For a metallic system such as graphene16, the gap between the p_eth and the p_e+1st eigenvalues is typically very small as can be observed from Figure 2.

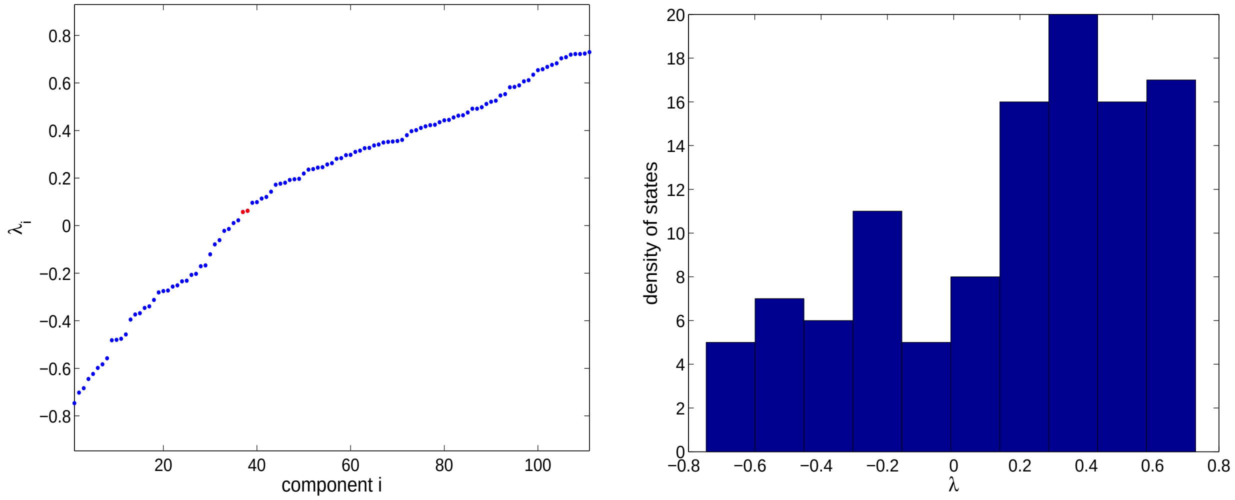


(a) eigenvalues

(b) Histogram of eigenvalues

FIG. 1. *The distribution of the first 45 eigenvalues of the benzene Hamiltonian at T = 1000K.*

In Figure 3, we show that the SCF iteration behaves quite differently when it is applied to the KSDFT and EDFT models for the metallic system graphene16. The temperature for the EDFT model is set to 1000K. All parameters for the SCF iteration such as the initial guess to the wavefunctions, the number of LOBPCG iterations used to solve the linear eigenvalue problem at each SCF iteration etc. are the same for both runs. We plot the residual norm $\|H(X)X - X(X^*HX)\|$ against the iteration number. We can clearly see from this figure that the SCF iteration converges steadily to a self-consistent solution for the EDFT model, whereas the residual norm associated with the standard KSDFT model stagnates around 10^{-3} after 60 iterations. This stagnation is caused by the ill-posedness of



(a) eigenvalues

(b) Histogram of eigenvalues

FIG. 2. *The distribution of the first 111 eigenvalues of the graphene16 Hamiltonian at T = 1000K.*

the standard KSDFT model for this particular metallic system.

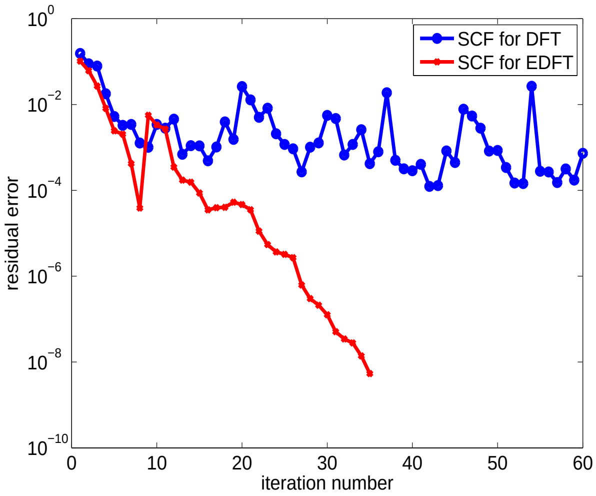


FIG. 3. *A comparion between the convergence behaviors of the SCF iterations when applied to the KSDFT and EDFT models for the graphene16 system.*

We now report the overall performance of SCF and NPG on the systems listed in Table 1. A summary of computational results for temperatures 50K, 300K, 1000K and 3000K is presented in Tables 2, 3, 4 and 5, respectively. In these tables, “iter” denotes the number of iterations used, “res” denotes the residual $\|HX - X(X^*HX)\|_F$ at the final iteration X, “feasi” denotes the violation of orthogonality constraints $\|X^*X - I\|_F$, and “time” denotes the runtime measured in seconds. From these tables, we can observe that the ensemble DFT model improves the convergence of metallic systems in terms of the iteration number and cpu time as the temperature increases in most cases for both NPG and SCF. These models do not change the convergence behavior of insulating systems. The performance of NPG is comparable to or better than SCF on achieving similar accuracy. Although NPG may be slower than SCF in terms of runtime on some cases of hnnco and ptnio, NPG is able to find a solution with lower total energy than that of SCF.

We should point out that the number of iterations in SCF and NPG should not be compared directly since the computational cost of each iteration of SCF is much expensive than that of NPG as $p \ll n$. Specifically, each iteration of NPG needs to compute the SVD of an $n \times p$ matrix Y. Let $V\Sigma V^*$ be the eigenvalue decomposition of Y^*Y. Then $YV\mathrm{Diag}(\Sigma_{11}^{-1}, \ldots, \Sigma_{pp}^{-1})$ and V are the left and right singular vectors of Y, respectively. On the other hand, each iteration of SCF needs to compute the smallest p eigenvalues and their corresponding eigenvectors of the $n \times n$ Hamiltonian matrix. When an iterative solver such as the LOBPCG algorithm is used to compute these eigenvectors, the most time consuming computation in each iteration is often the multiplication of the Hamiltonian with a vector. Thus, the overall cost of SCF should be measured in terms of the total number of SCF iterations times the number of iterations used to solve the linear eigenvalue problem.

Finally, the convergence history of the occupation numbers associated with the first p_e wavefunctions of the benzene and graphene16 Hamiltonian on T=300K and 1000K are depicted in Figures 4 and 5, respectively. Specifically, the occupation numbers of the benzene Hamiltonian are either very close to 0 or 1, while some occupation numbers of the graphene16 Hamiltonian are significantly larger than 0 but far away from 1. The relative reduction of the total energy defined by

$$\Delta\mathcal{M}(Z^k) = \mathcal{M}(Z^k) - \min\{\mathcal{M}(Z^i)\}$$

and the residual $\|HX^k - X^k((X^k)^*HX^k)\|_F$ on T=300K and 1000K are presented in Figures 6 and 7, respectively. Although both errors are not decreased monotonically, the trend of decreasing is clear.

TABLE 2
Computational results of EDFT using temperature 50K.

	SCF					NPG				
name	$\mathcal{M}(Z)$	iter	res	feasi	time	$\mathcal{M}(Z)$	iter	res	feasi	time
benzene	-3.7225751362901e+01	10	1.12e-06	1.62e-14	43.38	-3.7225750119414e+01	54	6.34e-06	1.84e-14	29.90
alanine	-6.1161921213037e+01	11	6.58e-06	1.79e-14	81.50	-6.1161920628696e+01	60	5.49e-06	2.37e-14	57.54
hnco	-2.8634664365489e+01	11	1.25e-06	1.04e-14	11.16	-2.8634648773849e+01	95	7.32e-06	8.94e-15	11.67
si2h4	-6.3009750459909e+00	10	8.29e-07	9.51e-15	8.56	-6.3009746553063e+00	53	6.86e-06	1.14e-14	6.33
ctube661	-1.3463843176500e+02	10	5.80e-06	4.07e-14	259.65	-1.3463836037853e+02	57	5.46e-06	3.59e-14	161.54
graphene16	-9.4046959319449e+01	54	8.14e-06	3.29e-14	380.92	-9.4046958128694e+01	262	7.59e-06	3.38e-14	174.12
ptnio	-2.2678880288682e+02	42	7.51e-06	3.69e-14	408.11	-2.2678861827323e+02	856	9.23e-06	3.20e-14	672.70
nic	-2.3543529955319e+01	9	1.95e-06	8.20e-15	3.93	-2.3543529531528e+01	46	8.17e-06	6.27e-15	3.28

TABLE 3
Computational results of EDFT using temperature 300K.

	SCF					NPG				
name	$\mathcal{M}(Z)$	iter	res	feasi	time	$\mathcal{M}(Z)$	iter	res	feasi	time
benzene	-3.7225751362901e+01	10	1.09e-06	1.78e-14	46.48	-3.7225750597810e+01	55	9.27e-06	1.67e-14	31.09
alanine	-6.1161921213037e+01	11	6.58e-06	1.79e-14	82.38	-6.1161920688748e+01	61	8.25e-06	2.34e-14	56.73
hnco	-2.8634664365396e+01	11	1.30e-06	1.11e-14	11.83	-2.8634657535018e+01	84	9.09e-06	8.49e-15	10.20
si2h4	-6.3009750459908e+00	10	8.85e-07	8.84e-15	7.91	-6.3009740148473e+00	52	9.61e-06	9.21e-15	5.78
ctube661	-1.3463843176499e+02	10	4.58e-06	4.08e-14	256.66	-1.3463836532271e+02	70	4.37e-06	3.26e-14	190.51
graphene16	-9.4048801256318e+01	37	9.65e-06	3.32e-14	248.57	-9.4048824827294e+01	214	8.43e-06	3.54e-14	145.94
ptnio	-2.2678904380333e+02	49	8.74e-06	3.67e-14	469.60	-2.2678972484319e+02	746	9.50e-06	3.23e-14	605.54
nic	-2.3543529955319e+01	9	1.95e-06	8.46e-15	4.64	-2.3543529557882e+01	50	4.04e-06	5.69e-15	3.11

5. Concluding remarks. The ensemble DFT model is a very important extension of the KS-DFT for metallic systems because it circumvents the numerical difficult associated with the potential absence of a spectrum gap in the vicinity of the Fermi level. The optimization formulation of the ensemble DFT is similar to that of KS-DFT except that the density and total energy functional is revised according to the fractional occupancies and some additional constraints on the occupancies are added. The special structure of the model enables us to eliminate the dependence on occupancies and establish an equivalent model with a single variable and a single spherical constraint. However, the projected gradient method on sphere cannot be applied directly since the gradient of the entropy term tends to be singular when some singular value of the new variable are close to 0 or 1. Inspired by the recent progress on sparse and

TABLE 4
Computational results of EDFT using temperature 1000K.

SCF						NPG				
name	$\mathcal{M}(Z)$	iter	res	feasi	time	$\mathcal{M}(Z)$	iter	res	feasi	time
benzene	-3.7225751239968e+01	10	7.01e-07	1.54e-14	44.83	-3.7225750733224e+01	66	1.19e-06	1.71e-14	34.03
alanine	-6.1161920827271e+01	11	6.53e-06	1.85e-14	87.00	-6.1161920541828e+01	62	3.40e-06	2.55e-14	63.02
hnco	-2.8634523058884e+01	11	2.04e-06	1.10e-14	11.31	-2.8634656221154e+01	117	4.80e-06	8.84e-15	14.83
si2h4	-6.3009735502548e+00	9	5.37e-06	8.02e-15	6.98	-6.3009709887415e+00	52	6.92e-06	9.51e-15	6.40
ctube661	-1.3463842783907e+02	10	5.97e-06	3.97e-14	270.57	-1.3463838680852e+02	62	7.26e-06	3.53e-14	180.16
graphene16	-9.4053900385484e+01	25	5.73e-06	3.37e-14	171.66	-9.4054235523318e+01	181	8.25e-06	3.33e-14	128.67
ptnio	-2.2679234640468e+02	38	4.95e-06	3.71e-14	371.35	-2.2679890365613e+02	474	9.51e-06	3.48e-14	388.96
nic	-2.3543529947006e+01	9	1.95e-06	8.07e-15	4.30	-2.3543528492570e+01	46	7.69e-06	7.26e-15	2.86

TABLE 5
Computational results of EDFT using temperature 3000K.

SCF						NPG				
name	$\mathcal{M}(Z)$	iter	res	feasi	time	$\mathcal{M}(Z)$	iter	res	feasi	time
benzene	-3.7222935075897e+01	10	8.31e-06	1.47e-14	43.84	-3.7225753284406e+01	76	3.50e-06	1.74e-14	39.44
alanine	-6.1158742483813e+01	13	2.99e-06	1.87e-14	98.89	-6.1161928693158e+01	98	8.15e-06	2.14e-14	81.97
hnco	-2.8629683282312e+01	11	7.49e-06	1.14e-14	10.33	-2.8635307352591e+01	208	8.59e-06	9.87e-15	24.17
si2h4	-6.2980107473949e+00	10	2.92e-06	7.50e-15	7.92	-6.3009896164398e+00	80	9.51e-06	8.21e-15	8.19
ctube661	-1.3463171186211e+02	11	8.32e-06	4.11e-14	284.09	-1.3463845201598e+02	79	7.31e-06	3.43e-14	221.22
graphene16	-9.4059107112515e+01	19	7.71e-06	3.27e-14	136.86	-9.4071570724660e+01	170	9.54e-06	3.38e-14	114.86
ptnio	-2.2682829690698e+02	28	6.73e-06	3.62e-14	273.64	-2.2686319823897e+02	304	8.27e-06	3.11e-14	254.87
nic	-2.3542691523733e+01	9	1.96e-06	8.45e-15	4.34	-2.3543508787591e+01	69	3.90e-06	6.63e-15	3.93

low-rank matrix optimization, we develop a proximal gradient method by keeping the entropy term and linearizing all other terms in the total energy functional. The proximal gradient subproblem is solved by estimating the Lagrangian multiplier of the spherical constraint. Convergence to a stationary point is established. The proximal gradient methods are further improved by the state-of-the-art acceleration techniques such as Barzilai-Borwein steps and non-monotone line search with global convergence guarantees. Numerical experiments show that our methods can be more efficient and robust than SCF on many metallic systems in the KSSOLV toolbox under the Matlab environment.

Acknowledgements. The authors are grateful to Prof. Eric Cances and two anonymous referees for their detailed and valuable comments and suggestions.

REFERENCES

[1] Arnaud Anantharaman and Eric Cancès, *Existence of minimizers for Kohn-Sham models in quantum chemistry*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 26 (2009), pp. 2425–2455.

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

[3] Jonathan Barzilai and Jonathan M. Borwein, *Two-point step size gradient methods*, IMA J. Numer. Anal., 8 (1988), pp. 141–148.

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

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

[6] Eric Cancès, *SCF algorithms for Hartree-Fock electronic calculations*, Lecture Notes in Chemistry, 74 (2000), pp. 17–43.

[7] ———, *Self-consistent field algorithms for Kohn-Sham models with fractional occupation numbers*, Journal of Chemical Physics, 114(24) (2001), p. 1061610622.

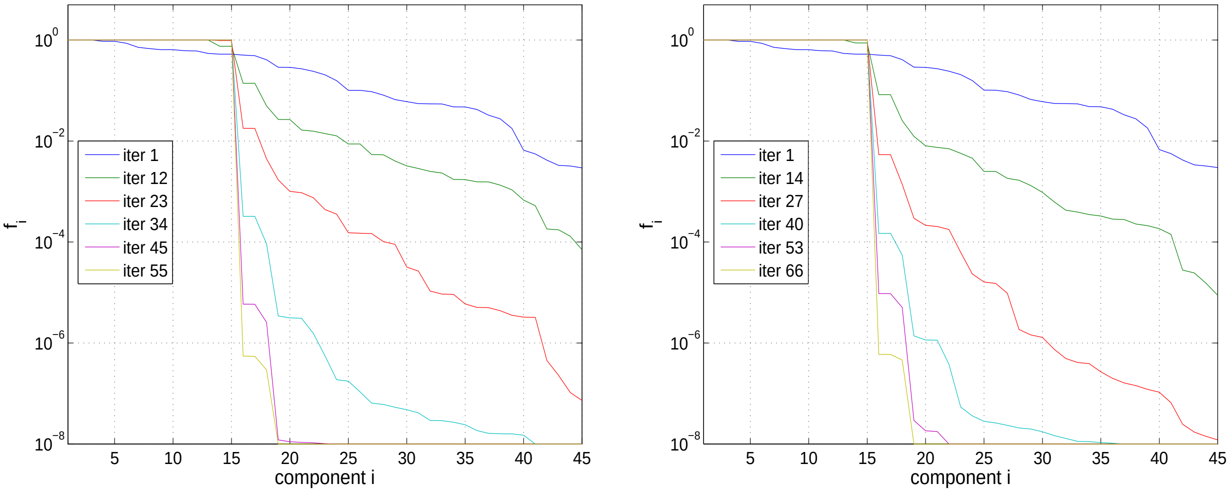
[8] Eric Cancès and Claude Le Bris, *Can we outperform the DIIS approach for electronic structure calculations?*, International Journal of Quantum Chemistry, 79(2) (2000), pp. 82–90.

[9] E. Cancès, M. Defranceschi, W. Kutzelnigg, C. Le Bris, and Y. Maday, *Handbook of numerical analysis. Volume X: special volume: computational chemistry*, North-Holland, 2003, ch. Computational quantum chemistry: a primer, pp. 3–270.

[10] Eric Cancès and Katarzyna Pernal, *Projected gradient algorithms for Hartree-Fock and density matrix functional theory calculations*, The Journal of Chemical Physics, 128 (2008).

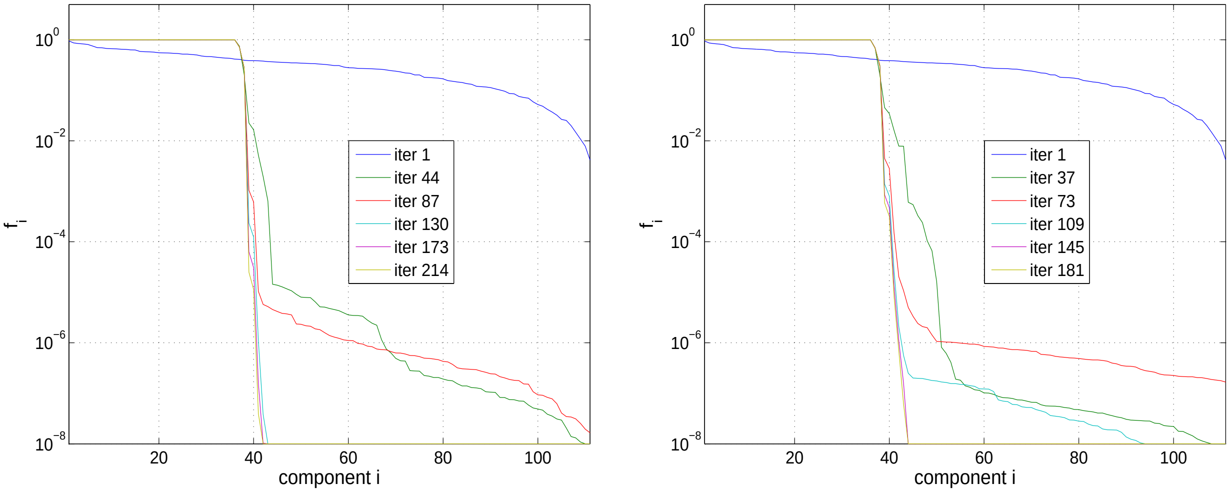
[11] Jeng-Da Chai, *Density functional theory with fractional orbital occupations*, The Journal of Chemical Physics, 136 (2012), pp. –.

[12] A. D. Daniels and G. E. Scuseria, *Converging difficult SCF cases with conjugate gradient density matrix search*, Phys. Chem. Chem. Phys., 2 (2000).



(a) T = 300K (b) T = 1000K

FIG. 4. *The convergence history of the occupation numbers associated with the first 45 wavefunctions of the benzene Hamiltonian.*



(a) T = 300K (b) T = 1000K

FIG. 5. *The convergence history of the occupation numbers associated with the first 111 wavefunctions of the graphene16 Hamiltonian.*

[13] CHAO DING, DEFENG SUN, JIE SUN, AND KIM-CHUAN TOH, *Spectral operators of matrices*, tech. report, arXiv:1401.2269, 2014.

[14] JULIANO B. FRANCISCO, JOSE MARIO MARTINEZ, AND LEANDRO MARTINEZ, *Globally convergent trust-region methods for self-consistent field electronic structure calculations*, The Journal of Chemical Physics, 121 (2004), pp. 10863–10878.

[15] JULIANO B. FRANCISCO, JOSÉ MARIO MARTÍNEZ, AND LEANDRO MARTÍNEZ, *Density-based globally convergent trust-region methods for self-consistent field electronic structure calculations*, J. Math. Chem., 40 (2006), pp. 349–377.

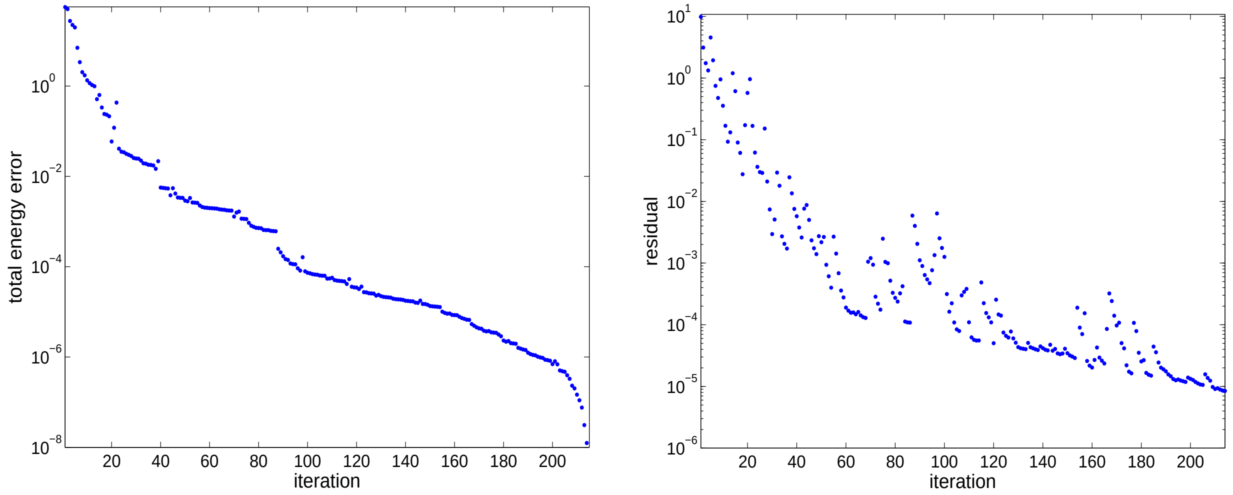
[16] RUPERT FRANK, ELLIOTT LIEB, ROBERT SEIRINGER, AND HEINZ SIEDENTOP, *Müller's exchange-correlation energy in density-matrix-functional theory*, Phys. Rev. A, 76 (2007), p. 052517.

[17] CHRISTOPH FREYSOLDT, SIXTEN BOECK, AND JÖRG NEUGEBAUER, *Direct minimization technique for metals in density functional theory*, Phys. Rev. B, 79 (2009), p. 241103.

[18] ELAINE T. HALE, WOTAO YIN, AND YIN ZHANG, *Fixed-point continuation for l_1-minimization: methodology and convergence*, SIAM J. Optim., 19 (2008), pp. 1107–1130.

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

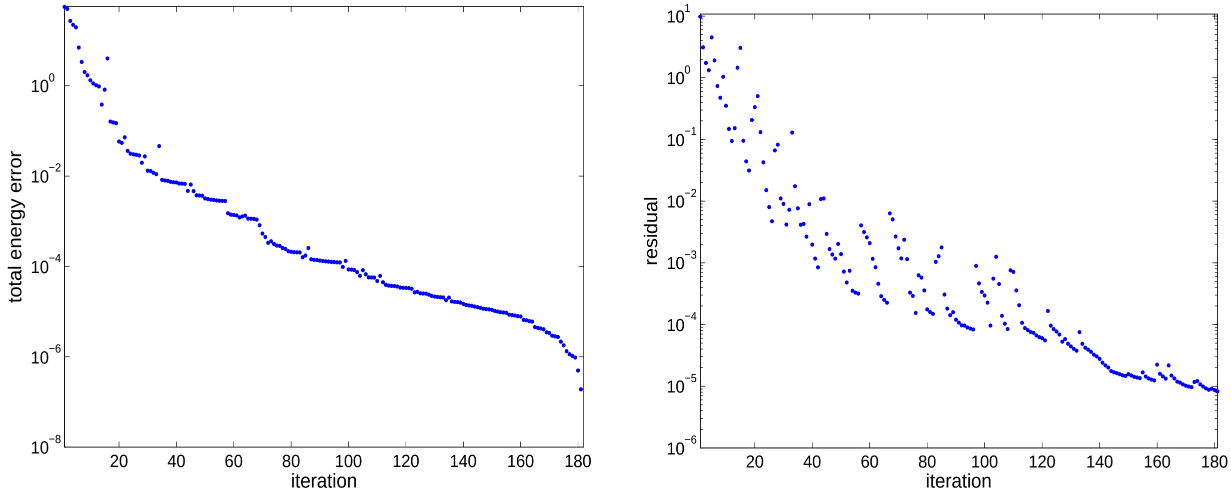
[20] ANDREW V. KNYAZEV, *Toward the optimal preconditioned eigensolver: locally optimal block preconditioned conjugate gradient method*,



(a) total energy error

(b) residual

FIG. 6. *The convergence history of the total energy errors and residuals of the graphene16 at T = 300K.*



(a) total energy error

(b) residual

FIG. 7. *The convergence history of the total energy errors and residuals of the graphene16 at T = 1000K.*

SIAM J. Sci. Comput., 23 (2001), pp. 517–541.

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

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

[23] G. KRESSE AND J. FURTHMLLER, *Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set*, Computational Materials Science, 6 (1996), pp. 15 – 50.

[24] A.S. LEWIS, *Derivatives of spectral functions*, Mathematics of Operations Research, 21 (1996).

[25] X.-P. LI, R. W. NUNES, AND DAVID VANDERBILT, *Density-matrix electronic-structure method with linear system-size scaling*, Phys. Rev. B, 47 (1993), pp. 10891–10894.

[26] L. LIN AND C. YANG, *Elliptic preconditioner for accelerating the self-consistent field iteration in Kohn-Sham density functional theory*, SIAM J. Sci. Comp, 35 (2013).

[27] X. LIU, Z. WEN, X. WANG, M. ULBRICH, AND Y. YUAN, *On the analysis of the discretized Kohn-Sham density functional theory*, tech. report, 2014. arXiv:1402.5052.

[28] Zhaosong Lu and Yong Zhang, *Iterative reweighted singular value minimization methods for l_p regularized unconstrained matrix minimization*, tech. report, Simon Fraser University, 2014.

[29] N. Marzari, D. Vanderbilt, and M.C. Payne, *Ensemble density-functional theory for ab initio molecular dynamics of metals and finite-temperature insulators*, Physical Review Letters, 79 (1997), pp. 1337–1340.

[30] N. David Mermin, *Thermal properties of the inhomogeneous electron gas*, Phys. Rev., 137 (1965), pp. A1441–A1443.

[31] John M. Millam and Gustavo E. Scuseria, *Linear scaling conjugate gradient density matrix search as an alternative to diagonalization for first principles electronic structure calculations*, The Journal of Chemical Physics, 106 (1997), pp. 5569–5577.

[32] M. C. Payne, M. P. Teter, D. C. Allan, T. A. Arias, and J. D. Joannopoulos, *Iterative minimization techniques for ab initio total-energy calculations: molecular dynamics and conjugate gradients*, Rev. Mod. Phys., 64 (1992), pp. 1045–1097.

[33] Bernd G Pfrommer, James Demmel, and Horst Simon, *Unconstrained energy functionals for electronic structure calculations*, Journal of Computational Physics, 150 (1999), pp. 287 – 298.

[34] E. Prodan and P. Nordlander, *On the Kohn-Sham equations with periodic background potentials*, Journal of Statistical Physics, 111 (2003).

[35] R.T. Rockafellar and R. J.-B. Wets, *Variatinoal Analysis*, Springer, third ed., 2009.

[36] Alvaro Ruiz-Serrano and Chris-Kriton Skylaris, *A variational method for density functional theory calculations on metallic systems with thousands of atoms*, The Journal of Chemical Physics, 139 (2013), pp. –.

[37] V. R. Saunders and I. H. Hillier, *A "Level-shifting" method for converging closed shell Hartree-Fock wave functions*, International Journal of Quantum Chemistry, 7 (1973), pp. 699–705.

[38] Lea Thogersen, Jeppe Olsen, Danny Yeager, Poul Jorgensen, Pawel Salek, and Trygve Helgaker, *The trust-region self-consistent field method: Towards a black-box optimization in Hartree–Fock and Kohn–Sham theories*, The Journal of Chemical Physics, 121 (2004), pp. 16–27.

[39] Troy Van Voorhis and Martin Head-Gordon, *A geometric approach to direct minimization*, Molecular Physics, 100 (2002), pp. 1713–1721.

[40] Joost VandeVondele and Jurg Hutter, *An efficient orbital transformation method for electronic structure calculations*, The Journal of Chemical Physics, 118 (2003), pp. 4365–4369.

[41] Z. Wen, A. Milzarek, M. Ulbrich, and H. Zhang, *Adaptive regularized self-consistent field iteration with exact Hessian for electronic structure calculation*, SIAM J. Sci. Comput, 35 (2013), pp. A1299–A1324.

[42] Zaiwen Wen and Wotao Yin, *A feasible method for optimization with orthogonality constraints*, Mathematical Programming, (2012).

[43] Zaiwen Wen, Wotao Yin, Donald Goldfarb, and Yin Zhang, *A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization and continuation*, SIAM Journal on Scientific Computing, 32 (2010), pp. 1832–1857.

[44] Zaiwen Wen, Wotao Yin, Hongchao Zhang, and Donald Goldfarb, *On the convergence of an active set method for l_1 minimization*, Optimization Methods and Software, 27 (2012), pp. 1127–1146.

[45] C. Yang, W. Gao, and J. C. Meza, *On the convergence of the self-consistent field iteration for a class of nonlinear eigenvalue problems*, SIAM. J. Matrix Anal. Appl., 30 (2009), pp. 1773–1788.

[46] Chao Yang, Juan C. Meza, Byounghak Lee, and Lin-Wang Wang, KSSOLV—*a MATLAB toolbox for solving the Kohn-Sham equations*, ACM Trans. Math. Softw., 36 (2009), pp. 1–35.

[47] C. Yang, J. C. Meza, and L. Wang, *A trust region direct constrained minimization algorithm for the Kohn-Sham equation*, SIAM Journal of Scientific Computing, 29 (2007), pp. 1854–1875.

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

[49] Hongchao Zhang and William W. Hager, *A nonmonotone line search technique and its application to unconstrained optimization*, SIAM J. Optim., 14 (2004), pp. 1043–1056.

[50] X. Zhang, J. Zhu, Z. Wen, and A. Zhou, *Gradient type optimization methods for electronic structure calculations*, SIAM Journal on Scientific Computing, 36 (2014), pp. C265–C289.

[51] Aihui Zhou, *An analysis of finite-dimensional approximations for the ground state solution of Bose-Einstein condensates*, Nonlinearity, 17 (2004), pp. 541–550.