Numer. Math. Theor. Meth. Appl.
doi: 10.4208/nmtma.OA-2022-0026

Vol. **16**, No. 1, pp. 1-25
February 2023

Convergent and Orthogonality Preserving Schemes for Approximating the Kohn-Sham Orbitals

Xiaoying Dai[1,2,*], Liwei Zhang[1,2] and Aihui Zhou[1,2]

[1] *LSEC, Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China*
[2] *School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China*

Received 13 February 2022; Accepted (in revised version) 10 July 2022

Abstract. To obtain convergent numerical approximations without using any orthogonalization operations is of great importance in electronic structure calculations. In this paper, we propose and analyze a class of iteration schemes for the discretized Kohn-Sham Density Functional Theory model, with which the iterative approximations are guaranteed to converge to the Kohn-Sham orbitals without any orthogonalization as long as the initial orbitals are orthogonal and the time step sizes are given properly. In addition, we present a feasible and efficient approach to get suitable time step sizes and report some numerical experiments to validate our theory.

AMS subject classifications: 37M15, 37M21, 65M12, 65N25, 81Q05

Key words: Gradient flow based model, density functional theory, orthogonality preserving scheme, convergence, temporal discretization.

1. Introduction

Electronic structure calculations play an important role in numerous fields such as quantum chemistry, materials science and drug design. Due to the good balance of accuracy and computational cost, the Kohn-Sham Density Function Theory (DFT) model [20,22,29,31,32] has become one of the most widely used models in electronic structure calculations which is usually treated as either a nonlinear eigenvalue problem (Kohn-Sham equation) or an orthogonality constraint minimization problem (Kohn-Sham total energy direct minimization problem).

[*]Corresponding author. *Email addresses:* zhanglw@lsec.cc.ac.cn (L. Zhang), azhou@lsec.cc.ac.cn (A. Zhou), daixy@lsec.cc.ac.cn (X. Dai)

In the literature, there are a number of works on the design and analysis of numerical methods for solving the Kohn-Sham equation (see, e.g., [5, 7, 8, 14, 24, 37, 44] and references cited therein). To obtain the solution of this nonlinear eigenvalue problem, we observe that some self consistent field (SCF) iterations are usually used [29] (see also [2, 4, 21, 25, 33, 34, 47]). Unfortunately, the convergence of SCF iterations is uncertain. We understand that its convergence has indeed been investigated when there is a sufficiently large gap between the occupied and unoccupied states and the second-order derivatives of the exchange correlation functional are uniformly bounded from above [3, 26, 27, 41], which is important in the theoretical point of view. It becomes significant to investigate the convergence of SCF iterations when the gap is not large in application.

We see that an alternative approach to obtain the ground states is to solve the Kohn-Sham total energy minimization problem, which is an orthogonality constrained minimization problem [32]. The direct minimization approach attracts the attention of many researchers in recent years [6, 17, 30], and many different kinds of optimization methods are applied to electronic structure calculations and investigated (see, e.g., [10, 11, 18, 19, 38, 42, 45, 46]).

For solving either the nonlinear eigenvalue problem or the orthogonality constrained minimization problem, except for few works such as [19], the orthogonalization procedure is usually required, which is very expensive and limits the parallel scalability in numerical implementation.

Recently, Dai *et al.* proposed a gradient flow based Kohn-Sham DFT model [12] that is a time evolution problem and is completely different from either the nonlinear eigenvalue problem or the orthogonality constrained minimization problem. It is proved in [12] that the flow of the new model is orthogonality preserving, and the solution can evolve to the ground state. Consequently, the gradient flow based model provides a novel and attractive approach for solving Kohn-Sham DFT apart from the eigenvalue problem model and the energy minimization model. In other words, the gradient flow based model is quite promising in ground state electronic structure calculations and deserves further investigation. For the sake of clarity, we would like to mention that the gradient flow based Kohn-Sham DFT model is different from the time dependent Kohn-Sham equation in [28, 36, 43].

In this paper, we propose a general framework of orthogonality preserving schemes that produce efficient approximations of the Kohn-Sham orbitals with the help of the gradient flow based model. In addition, we prove the global convergence and local convergence rate of the new schemes under some mild assumptions. We also provide some typical choices for the auxiliary mapping appeared in the framework, and a feasible and efficient approach to obtain the desired time step sizes that satisfy the assumptions required in the analysis, which result in several typical orthogonality preserving schemes that can produce convergent approximations of the Kohn-Sham orbitals.

The rest of the paper is organized as follows. In Section 2, we briefly review the gradient flow based Kohn-Sham DFT model and some notation frequently used throughout this paper. We then propose a framework for orthogonality preserving schemes for solving the discretized Kohn-Sham model in Section 3 and prove its global convergence as well as local convergence rate under some reasonable assumptions and with proper time step size. Then, in Section 4, we provide some specific choices for the auxiliary mapping and the time step size. We then report some numerical results obtained by the proposed schemes in Section 5 to verify our theory. Finally, we give some concluding remarks in Section 6.

2. Kohn-Sham DFT models

2.1. Classical Kohn-Sham DFT model

According to the Kohn-Sham density functional theory [22], the ground state of a system can be obtained by solving

$$\begin{aligned} &\inf_{U\in (H^1(\mathbb{R}^3))^N} E_{\text{KS}}(U) \\ &\text{s.t.} \quad U^T U = I_N, \end{aligned} \tag{2.1}$$

where

$$\begin{aligned} &U = (u_1, \ldots, u_N) \in \left(H^1(\mathbb{R}^3)\right)^N, \\ &U^T V = \left(\langle u_i, v_j\rangle_{L^2(\mathbb{R}^3)}\right)_{i,j=1}^N, \quad \forall U, V \in \left(H^1(\mathbb{R}^3)\right)^N, \end{aligned}$$

and the objective functional $E_{\text{KS}}(U)$ reads as

$$\begin{aligned} E_{\text{KS}}(U) = &\frac{1}{2}\int_{\mathbb{R}^3} \sum_{i=1}^N |\nabla u_i(r)|^2 dr + \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\rho(r)\rho(r')}{|r-r'|} dr dr' \\ &+ \int_{\mathbb{R}^3} V_{ext}(r)\rho(r)dr + \int_{\mathbb{R}^3} \varepsilon_{xc}(\rho)(r)\rho(r)dr. \end{aligned} \tag{2.2}$$

Here, N denotes the number of electrons, $\{u_i\}_{i=1,2,\ldots,N}$ are usually called the Kohn-Sham orbitals, $\rho(r) = \sum_{i=1}^N |u_i(r)|^2$ is the electronic density (we assume each Kohn-Sham orbital is occupied by one electron here), $V_{ext}(r)$ is the external potential generated by the nuclei, and $\varepsilon_{xc}(\rho)(r)$ is the exchange-correlation functional which is not known explicitly. In practice, some approximation such as local density approximation (LDA), generalized gradient approximation (GGA) or some other approximations has to be used [29].

We see that the feasible set of (2.1) is a Stiefel manifold which is defined as

$$\mathcal{M}^N = \left\{U \in \left(H^1(\mathbb{R}^3)\right)^N : U^T U = I_N\right\}. \tag{2.3}$$

To get rid of the nonuniqueness of the minimizer caused by the invariance of the energy functional under orthogonal transformations to the Kohn-Sham orbitals (i.e., $E(U) = E(UP), \forall P \in \mathcal{O}^N$ with $\mathcal{O}^N$ being the set of orthogonal matrices of order N), we, following [10, 12, 13], consider (2.1) on the Grassmann manifold $\mathcal{G}^N$ which is a quotient manifold of Stifel manifold, that is

$$\mathcal{G}^N = \mathcal{M}^N / \sim .$$

Here, $\sim$ denotes the equivalence relation which is defined as: $\hat{U} \sim U$ if and only if there exists $P \in \mathcal{O}^N$ such that $\hat{U} = UP$. For any $U \in \mathcal{M}^N$, we denote

$$[U] = \left\{ UP : P \in \mathcal{O}^N \right\},$$

then the Grassmann manifold $\mathcal{G}^N$ can be formulated as

$$\mathcal{G}^N = \left\{ [U] : U \in \mathcal{M}^N \right\}.$$

For $[U] \in \mathcal{G}^N$, the tangent space of $[U]$ on $\mathcal{G}^N$ is the following set:

$$\mathcal{T}_{[U]}\mathcal{G}^N = \left\{ W \in V^N : W^T U = \mathbf{0} \in \mathbb{R}^{N\times N} \right\}. \tag{2.4}$$

In this paper, we assume that (2.1) achieves its minimum in $\mathcal{G}^N$, which implies that (2.1) is equivalent to

$$\min_{[U]\in\mathcal{G}^N} E(U). \tag{2.5}$$

In addition, we see from [1] that the Grassmann gradient of $E_{\text{KS}}(U)$ is

$$\nabla_G E_{\text{KS}}(U) = \nabla E_{\text{KS}}(U) - UU^T \nabla E_{\text{KS}}(U),$$

where

$$\nabla E_{\text{KS}}(U) = \mathcal{H}(\rho)U, \quad \forall\, U \in \left(H^1(\mathbb{R}^3)\right)^N$$

is the Euclidean gradient of $E_{\text{KS}}(U)$,

$$\mathcal{H}(\rho) = -\frac{1}{2}\Delta + V_{ext} + \int_{\mathbb{R}^3} \frac{\rho(r')}{|r-r'|} dr' + v_{xc}(\rho)$$

is symmetric and

$$v_{xc}(\rho) = \frac{\delta\big(\rho\varepsilon_{xc}(\rho)\big)}{\delta\rho}.$$

For any $U \in \mathcal{M}^N$, similar to [12], we may write $\nabla_G E_{KS}(U)$ as $\mathcal{A}_U U$, that is,

$$\nabla_G E_{\text{KS}}(U) = \mathcal{A}_U U,$$

where

$$\mathcal{A}_U = \nabla E_{\text{KS}}(U) U^T - U \nabla E_{\text{KS}}(U)^T$$

is anti-symmetric, i.e., ${\mathcal{A}_U}^T = -\mathcal{A}_U$. Furthermore, the Hessian of $E_{\text{KS}}(U)$ on $\mathcal{G}_N$ has the form [10]

$$\nabla_G^2 E(U)[D_1, D_2] = \text{tr}\big({D_1}^T \mathcal{H}(\rho) D_2\big) - \text{tr}\big({D_1}^T D_2 U^T \mathcal{H}(\rho) U\big)$$

$$+ 2\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{\big(\sum_i u_i(r)d_{1,i}(r)\big)\big(\sum_j u_j(r')d_{2,j}(r')\big)}{|r-r'|}drdr'$$
$$+ 2\int_{\mathbb{R}^3}\frac{\delta^2\big(\varepsilon_{xc}(\rho)\rho\big)}{\delta\rho^2}(r)\Big(\sum_i u_i(r)d_{1,i}(r)\Big)\Big(\sum_j u_j(r)d_{2,j}(r)\Big)dr$$

provided that the total energy functional is twice differentiable, or more specifically, the approximated exchange-correlation functional is twice differentiable. Here,

$$D_i = (d_{i,1}, d_{i,2}, \cdots, d_{i,N}) \in \mathcal{T}_{[U]}\mathcal{G}^N, \quad i = 1, 2.$$

The ground state of a system can also be obtained by considering the Euler-Lagrange equation of (2.1), which reads as

$$\begin{cases}\mathcal{H}(\rho)U = U\Lambda,\\ U \in \mathcal{M}^N.\end{cases} \tag{2.6}$$

The nonlinear eigenvalue problem (2.6) is indeed the so-called Kohn-Sham equation. For decades, the Kohn-Sham DFT models are investigated as either a minimization problem (2.5) or an eigenvalue problem (2.6).

In practice, we may discretize the Kohn-Sham energy minimization model (2.5) as well as the nonlinear eigenvalue model (2.6) by, e.g., the plane wave method, the local basis set method, or real space methods. More details about the discretization methods can be found in, for instance, the review paper [37]. If we choose a N_g-dimension space $V_{N_g} \subset H^1(\mathbb{R}^3)$ to approximate $H^1(\mathbb{R}^3)$, then the associated discretized Kohn-Sham model can be formulated as

$$\min_{[U]\in\mathcal{G}^N_{N_g}} E_{KS}(U), \tag{2.7}$$

or

$$\begin{cases}\mathcal{H}(\rho)U = U\Lambda,\\ U \in \mathcal{M}^N_{N_g},\end{cases} \tag{2.8}$$

where $\mathcal{G}^N_{N_g}$ is the discretized Grassmann manifold defined by

$$\mathcal{G}^N_{N_g} = \mathcal{M}^N_{N_g}/\sim$$

and

$$\mathcal{M}^N_{N_g} = \{U \in (V_{N_g})^N : U^TU = I_N\}$$

is the discretized Stiefel manifold with the equivalence relation $\sim$ having the similar meaning to what we have mentioned. Usually, $N_g \gg N$. We should point out that the operators on the discretized manifold, such as the Grassmann gradient and the Grassmann Hessian, have exactly the same forms as those on the continuous manifold.

2.2. Gradient flow based Kohn-Sham DFT model

Different from the minimization model (2.7) and the eigenvalue model (2.8), a gradient flow based Kohn-Sham DFT model was proposed in [12], which has the following form:

$$\begin{cases}\dfrac{d}{dt}U(t) = -\nabla_G E\big(U(t)\big), & \forall t \in \mathbb{R}^+,\\ U(0) = U_0 \in \mathcal{M}^N_{N_g}.\end{cases} \tag{2.9}$$

Here, as U_0 is required to be orthogonal, we see from [12] that for the gradient flow based model (2.9), there hold

$$U(t) \in \mathcal{M}^N_{N_g}, \quad \forall t \geq 0 \tag{2.10}$$

and

$$\frac{d}{dt}E\big(U(t)\big) = -\big\|\nabla_G E\big(U(t)\big)\big\|^2 \leq 0, \quad \forall t \geq 0. \tag{2.11}$$

Besides, it was proved in [12] that the norm of the extended gradient of energy functional exponentially decays to zero over time t, and the solution $U(t)$ will evolve to the ground state under some mild assumptions. We mention that the detailed derivation of the gradient flow based Kohn-Sham model can be found in [12].

3. A general framework of orthogonality preserving schemes

With the help of the gradient flow based model (2.9), we develop a general framework that enables us to obtain a class of orthogonality preserving schemes for getting convergent approximations of Kohn-Sham orbitals. To propose our numerical schemes, we first introduce the partition of the time interval

$$0 = t_0 < t_1 < \cdots < t_n < \cdots .$$

In addition, we denote $\Delta t_n = t_{n+1} - t_n$ and use U_n to symbolize the approximation of $U(t_n)$ for $n \in \mathbb{N}_0$ where $\mathbb{N}_0$ is the set of nonnegative integers.

3.1. Scheme framework

Given $U_0 \in \mathcal{M}^N_{N_g}$, we consider the following recursive scheme on the interval $[t_n, t_{n+1})$:

$$\begin{cases}\tilde{U}(t) - U_n = -(t-t_n)\mathcal{A}_{U^{\mathrm{Aux}}(t)}\dfrac{U_n + \tilde{U}(t)}{2}, & t \in [t_n, t_{n+1}),\\ U_{n+1} = \tilde{U}(t^-_{n+1}).\end{cases} \tag{3.1}$$

Here $U^{\mathrm{Aux}} : \mathbb{R} \to (V_{N_g})^N$ is an auxiliary piecewise smooth mapping which satisfies the interpolation condition $U^{\mathrm{Aux}}(t_n) = U_n$ for all n. We hence name our schemes as interpolation based schemes. Note that this interpolation condition is the only restriction on

U^{Aux}, which makes our framework quite flexible. As a result, we obtain the following framework for interpolation based scheme (Algorithm 1) for solving (2.9).

Algorithm 1: A framework for interpolation based scheme.

```
Given $\epsilon > 0$, initial orbitals $U_0 \in \mathcal{M}^N_{N_g}$, calculate the gradient $\nabla_G E(U_0)$ and let
    $n = 0$, $t_0 = 0$.
while $\|\nabla_G E(U_n)\| > \epsilon$ do
    Choose a suitable $\Delta t_n > 0$ and let $t_{n+1} = t_n + \Delta t_n$.
    Define $U^{\mathrm{Aux}}(t), t \in [t_n, t_{n+1})$ such that $U^{\mathrm{Aux}}(t_n) = U_n$.
    Update $U_{n+1} = \lim_{t \to t^-_{n+1}} \tilde{U}(t)$ with $\tilde{U}(t)$ satisfying
        $$\tilde{U}(t) - U_n = -(t - t_n)\mathcal{A}_{U^{\mathrm{Aux}}(t)} \frac{U_n + \tilde{U}(t)}{2}, \quad t \in [t_n, t_{n+1}). \qquad (3.2)$$
    Let $n = n + 1$, and calculate the gradient $\nabla_G E(U_n)$.
```

We see from Algorithm 1 that the time step sizes in our scheme can be provided step by step and adaptively, i.e., we may make a full use of the information obtained during the iteration to determine a suitable time step size at each step. Besides, we point out here that if the definition of $U^{\mathrm{Aux}}(t), t \in [t_n, t_{n+1})$ is independent of $\tilde{U}(t), t \in [t_n, t_{n+1})$ at the n-th iteration, then the problem (3.2) is said to be linear, and the corresponding scheme is called an explicit scheme. Otherwise, it is an implicit scheme. The following Theorem 3.1 shows that Algorithm 1 preserves the orthogonality of iterations no matter whether it is explicit or not.

Theorem 3.1. *If $\{U_n\}_{n\in\mathbb{N}_0}$ is produced by Algorithm 1, then $\{U_n\}_{n\in\mathbb{N}_0} \subset \mathcal{M}^N_{N_g}$.*

Proof. By rearranging (3.1), we have that

$$U_{n+1} = \left(I + \frac{\Delta t_n}{2}\mathcal{A}_{U^{\mathrm{Aux}}(t^-_{n+1})}\right)^{-1}\left(I - \frac{\Delta t_n}{2}\mathcal{A}_{U^{\mathrm{Aux}}(t^-_{n+1})}\right)U_n. \tag{3.3}$$

Since $\mathcal{A}_{U^{\mathrm{Aux}}(t^-_{n+1})}$ is anti-symmetric, we see that

$$\left(I + \frac{\Delta t_n}{2}\mathcal{A}_{U^{\mathrm{Aux}}(t^-_{n+1})}\right)^{-1}\left(I - \frac{\Delta t_n}{2}\mathcal{A}_{U^{\mathrm{Aux}}(t^-_{n+1})}\right)$$

forms a Cayley transformation. Hence, $U_{n+1} \in \mathcal{M}^N_{N_g}$ as long as $U_n \in \mathcal{M}^N_{N_g}$. Note that $U_0 \in \mathcal{M}^N_{N_g}$, we complete the proof by induction. □

In fact, we see from the proof of Theorem 3.1 that $\tilde{U}(t)$ is orthogonal for all $t \in \mathbb{R}^+$, namely, $\tilde{U}(t) \subset \mathcal{M}^N_{N_g}$. In addition, we see that for any explicit scheme, the orbitals U_{n+1} can be updated simply by (3.3). Therefore, to update U_{n+1} at each iteration of an explicit scheme, the main cost is to compute the inverse of

$$I + \frac{\Delta t_n}{2}\mathcal{A}_{U^{\mathrm{Aux}}(t^-_{n+1})},$$

which is a N_g-dimensional matrix inverse problem and is very expensive to obtain. Even though we can deal with it by solving the corresponding linear system using some iterative methods, it is still not cheap, especially when N_g is large.

Fortunately, we observe that $\mathcal{A}_U$ $(\forall U \in V^N)$ is anti-symmetric and has the following factorization:

$$\mathcal{A}_U = \begin{pmatrix} \nabla E(U) & U \end{pmatrix} \begin{pmatrix} U^T \\ -\nabla E(U)^T \end{pmatrix}.$$

Hence, by applying the Sherman-Morrison-Woodbury (SMW) formula [12,15], we have

$$\begin{aligned} \left(I + \frac{\Delta t_n}{2}\mathcal{A}_U\right)^{-1} &= I - \frac{\Delta t_n}{2}\begin{pmatrix} \nabla E(U) & U \end{pmatrix} \\ &\quad \times \left[I_{2N} + \frac{\Delta t_n}{2} \begin{pmatrix} U^T \nabla E(U) & U^T U \\ -\nabla E(U)^T \nabla E(U) & -\nabla E(U)^T U \end{pmatrix} \right]^{-1} \\ &\quad \times \begin{pmatrix} U^T \\ -\nabla E(U)^T \end{pmatrix}, \end{aligned} \tag{3.4}$$

which reduces the dimension of the matrix inverse problem significantly from N_g to $2N$. Therefore, we only need to deal with a linear system of dimension $2N$.

3.2. Numerical analysis

Note that (2.11) indicates the energy functional is non-increasing with respect to t, we may impose the following assumption on the time step sizes to maintain a similar property. We will show the existence of the desired time partition and introduce an efficient strategy to obtain such time step sizes in the next section.

Assumption 3.1. The sequence $\{t_n\}_{n=0}^{\infty}$ satisfies

$$\sum_{n=0}^{\infty} \Delta t_n = +\infty, \quad \text{i.e.,} \quad \lim_{n\to\infty} t_n = +\infty \tag{3.5}$$

and

$$\begin{aligned} E(U_{n+1}) - E(U_n) &= E\big(\tilde{U}(t_{n+1}^-)\big) - E(U_n) \\ &\leq -\eta \Delta t_n \|\nabla_G E(U_n)\|^2, \quad n \in \mathbb{N}_0 \end{aligned} \tag{3.6}$$

with $\eta > 0$ being a given parameter.

The condition (3.5) in Assumption 3.1 is simple and reasonable, as we are discretizing an infinite time period. Meanwhile, the condition (3.6) follows from (2.11), which indicates that the finite difference approximation of the temporal derivative stated in the left-hand side of (2.11) is somewhat comparable to $\|\nabla_G E(U(t))\|^2$.

Under Assumption 3.1, we obtain the following asymptotic behaviour for the approximated solution of (2.9) produced by Algorithm 1.

Theorem 3.2. *If the sequence* $\{t_n\}_{n\in\mathbb{N}_0}$ *satisfies Assumption* 3.1, *then for the sequence* $\{U_n\}_{n\in\mathbb{N}_0}$ *produced by Algorithm* 1 *with an initial guess* $U_0 \in \mathcal{M}^N_{N_g}$, *there holds*

$$\liminf_{n\to\infty} \|\nabla_G E(U_n)\| = 0.$$

Proof. We see from Assumption 3.1 that

$$E(U_n) - E(U_{n+1}) \geq \eta \Delta t_n \|\nabla_G E(U_n)\|^2.$$

Hence,

$$E(U_0) - E_{\min} \geq \sum_{n=0}^{\infty} \big(E(U_n) - E(U_{n+1})\big) \geq \eta \sum_{n=0}^{\infty} \Delta t_n \|\nabla_G E(U_n)\|^2,$$

where $E_{\min}$ is the minimum of the energy functional $E(U)$. Thus,

$$\sum_{n=0}^{\infty} \Delta t_n \|\nabla_G E(U_n)\|^2 < \infty. \tag{3.7}$$

If $\liminf_{n\to\infty} \|\nabla_G E(U_n)\| > 0$, then there exists $\epsilon_0 > 0$ such that

$$\|\nabla_G E(U_n)\| \geq \epsilon_0, \quad \forall n \in \mathbb{N}_0.$$

Hence, we obtain from (3.5) that

$$\sum_{n=0}^{\infty} \Delta t_n \|\nabla_G E(U_n)\|^2 \geq \epsilon_0^2 \sum_{n=0}^{\infty} \Delta t_n = \infty,$$

which is contradictory to (3.7). As a result,

$$\liminf_{n\to\infty} \|\nabla_G E(U_n)\| = 0.$$

The proof is complete. □

We see that a sufficient condition for (3.5) is $\Delta t_n > \tau, \forall n \in \mathbb{N}_0$ for some $\tau > 0$. Under this setting, Theorem 3.2 indicates that the sequence $\{U_n\}_{n=0}^{\infty}$ produced by Algorithm 1 will converge to an equilibrium point of (2.9) (at least for a subsequence). If the equilibrium point (denoted by U^*) is a local minimizer of the Kohn-Sham energy functional $E(U)$, we assume in addition that the Hessian of the Kohn-Sham energy functional is bounded from both above and below in a neighborhood of $[U^*]$, that is to say, the following assumption holds.

Assumption 3.2. There exists $\delta_1 > 0$, such that for all $[U] \in B([U^*], \delta_1)$,

$$\nabla_G^2 E(U)[D,D] \geq \underline{c}\|D\|^2, \quad \forall D \in \mathcal{T}_{[U]}\mathcal{G}^N_{N_g}, \tag{3.8}$$

$$\big\|\nabla_G^2 E(U)[D]\big\| \leq \bar{c}\|D\|, \quad \forall D \in \mathcal{T}_{[U]}\mathcal{G}^N_{N_g}, \tag{3.9}$$

where U^* is the local minimizer of $E(U)$ and $\bar{c} \geq \underline{c} > 0$ are some constants. Here, $B([U],\delta)$ is defined as

$$B\big([U],\delta\big) := \Big\{[V] \in \mathcal{G}^N_{N_g} : \min_{P\in\mathcal{O}^{N\times N}} \|U - VP\| \leq \delta\Big\}.$$

We mention that the positiveness condition (3.8) has been justified and used in, e.g., [10, 12, 38], which is related to the spectral gap of Hamiltonian. Meanwhile, the boundedness condition (3.9) is quite natural as it holds true for any fixed $[U]$ with some constants $\bar{c}_{[U]} > 0$, and we just require that there is a uniform upper bound $\bar{c}$ for all $\{\bar{c}_{[U]}\}_{[U]\in B([U^*],\delta_1)}$.

Remark 3.1. Under Assumption 3.2, there exists a positive constant C such that

$$\|\nabla_G E(U)\| \le C, \quad \forall U \in \mathcal{M}^N_{N_g},$$

where C can be chosen as $\sqrt{2}\bar{c}N$.

Besides, we review some preliminaries on the Grassmann manifold which will be used in the following analysis. Let $[U],[W] \in \mathcal{G}^N_{N_g}$, with $U, W \in \mathcal{M}^N_{N_g}$. We obtain from [10, Lemma A.1] that there exists a geodesic

$$\Gamma(t) = \left[UA\cos(\Theta t)A^T + A_2\sin(\Theta t)A^T\right], \quad t \in [0,1], \tag{3.10}$$

such that

$$\Gamma(0) = [U], \quad \Gamma(1) = [W].$$

Here,

$$U^TW = A\cos\Theta B^T, \quad W - U(U^TW) = A_2\sin\Theta B^T$$

is the SVD of U^TW and $W - U(U^TW)$, respectively,

$$\Theta = \mathrm{diag}(\theta_1, \theta_2, \cdots, \theta_N)$$

is a diagonal matrix with $\theta_i \in [0, \pi/2]$ and

$$\sin(\Theta t) = \mathrm{diag}\left(\sin(\theta_1 t), \sin(\theta_2 t), \cdots, \sin(\theta_N t)\right)$$

with a similar notation for $\cos(\Theta t)$. Note that $A_2 \in \mathcal{M}^N_{N_g}$.

Remark 3.2. For any $U \in \mathcal{M}^N, D \in \mathcal{T}_{[U]}\mathcal{G}^N$, let $D = ASB^T$ be the SVD of D where $A \in \mathcal{T}_{[U]}\mathcal{G}^N_{N_g}$, $S, B \in \mathbb{R}^{N\times N}$, then there exists an unique geodesic

$$\Gamma(t) = \left[UB\cos(St)B^T + A\sin(St)B^T\right], \tag{3.11}$$

which starts from $[U]$ and proceeds with direction D [16]. The above expression (3.10) is just a special case with direction $D = A_2\Theta A^T$.

More specifically, we use macro $[\exp_{[U]}(tD)]$ to denote the geodesic on $\mathcal{G}^N_{N_g}$ which starts with $[U]$ and proceeds with the initial direction $D \in \mathcal{T}_{[U]}\mathcal{G}^N_{N_g}$. We now define the parallel mapping which maps a tangent vector along the geodesic [16].

Definition 3.1. *The parallel mapping* $\tau_{(U,D,t)} : \mathcal{T}_{[U]}\mathcal{G}_{N_g}^N \to \mathcal{T}_{[\exp_{[U]}(tD)]}\mathcal{G}_{N_g}^N$ *along the geodesic* $[\exp_{[U]}(tD)]$ *is defined as*

$$\tau_{(U,D,t)}\tilde{D} = \big((-U\sin{(St)} + A\cos{(St)}A^T + (I_N - AA^T)\big)\tilde{D},$$

where $D = ASB^T$ *is the SVD of* D.

It can be verified that

$$\|\tau_{(U,D,t)}\tilde{D}\| = \|\tilde{D}\|, \quad \forall \tilde{D} \in \mathcal{T}_{[U]}\mathcal{G}_{N_g}^N. \tag{3.12}$$

To state our theory, we introduce two distances on the Grassmann manifold $\mathcal{G}_{N_g}^N$

$$\begin{aligned} &\mathrm{dist}_{cF}\big([U],[W]\big) = \min_{P\in\mathcal{O}^{N\times N}} \|U - WP\|, \\ &\mathrm{dist}_{geo}\big([U],[W]\big) = \|A_2\Theta A^T\|. \end{aligned} \tag{3.13}$$

Remark 3.3. It can be calculated that [16]

$$\begin{aligned} &\mathrm{dist}_{cF}\big([U],[W]\big) = \|2\sin{(\Theta/2)}\|, \\ &\mathrm{dist}_{geo}\big([U],[W]\big) = \|\Theta\|, \end{aligned}$$

which indicate that these two kinds of distance are equivalent. More specifically,

$$\mathrm{dist}_{cF}\big([U],[W]\big) \le \mathrm{dist}_{geo}\big([U],[W]\big) \le 2\,\mathrm{dist}_{cF}\big([U],[W]\big).$$

In addition, we see that

$$\|D\| = \|A_2\Theta A^T\| = \|\Theta\|_F = \mathrm{dist}_{geo}\big([U],[W]\big), \tag{3.14}$$

where D is the initial direction of the geodesic (3.10).

Furthermore, we need the following conclusion, which can be obtained from [39, Remarks 3.2 and 4.2].

Proposition 3.1. *Suppose* $E(U)$ *is of second order differentiable, then for all* $U \in \mathcal{M}_{N_g}^N$, $D \in \mathcal{T}_{[U]}\mathcal{G}_{N_g}^N$, *there exists a* $\xi \in (0,t)$ *such that*

$$\begin{aligned} E\big(\exp_{[U]}(tD)\big) &= E(U) + t\big\langle \nabla_G E\big(\exp_{[U]}(\xi D)\big), \tau_{(U,D,\xi)}D\big\rangle \\ &= E(U) + t\big\langle \nabla_G E(U), D\big\rangle + \frac{t^2}{2}\nabla_G^2 E(U)[D,D] + o\left(t^2\|D\|^2\right), \end{aligned} \tag{3.15}$$

and

$$\tau_{(U,D,t)}^{-1}\nabla_G E\big(\exp_{[U]}(tD)\big) = \nabla_G E(U) + t\tau_{(U,D,\xi)}^{-1}\nabla_G^2 E\big(\exp_{[U]}(\xi D)\big)\left[\tau_{(U,D,\xi)}D\right].$$

Now we are ready to have the local convergence rate of the numerical approximations $\{U_n\}_{n=0}^{\infty}$ as stated in the following theorem.

Theorem 3.3. *Let Assumptions* 3.1 *and* 3.2 *hold true and assume that there exists a* $\tau > 0$ *such that* $\Delta t_n > \tau, \forall n \in \mathbb{N}_0$. *Then for the sequence* $\{U_n\}_{n\in\mathbb{N}_0}$ *produced by Algorithm* 1 *with an initial guess* $[U_0] \in B([U^*], \delta_1) \subset \mathcal{G}^N_{N_g}$, *there exists a constant* $\nu \in (0,1)$ *such that*

$$E(U_{n+1}) - E(U^*) \le \nu\big(E(U_n) - E(U^*)\big),$$

and hence, there exist $C_1, C_2 > 0$ *such that*

$$E(U_n) - E(U^*) \le C_1 \nu^n \operatorname{dist}_{geo}\big([U_0], [U^*]\big)^2,$$

and

$$\operatorname{dist}_{geo}\big([U_n], [U^*]\big) \le C_2 (\sqrt{\nu})^n \operatorname{dist}_{geo}\big([U_0], [U^*]\big).$$

Proof. For simplicity, we denote $d_n = \operatorname{dist}_{geo}([U_n], [U^*])$. We see that

$$\begin{aligned} E(U_{n+1}) - E(U^*) &= E(U_{n+1}) - E(U_n) + E(U_n) - E(U^*) \\ &\le -\eta \Delta t_n \|\nabla_G E(U_n)\|^2 + E(U_n) - E(U^*). \end{aligned}$$

For U_n and $U^* \in \mathcal{M}^N_{N_g}$, there exists a unique geodesic $\exp_{[U^*]}(tD_n)$ such that

$$\exp_{[U^*]}(0) = [U^*], \quad \exp_{[U^*]}(D_n) = [U_n],$$

where 0 is the zero element on the tangent space $\mathcal{T}_{[U^*]}\mathcal{G}^N_{N_g}$. Furthermore, there holds $\|D_n\| = d_n$.

We obtain from (3.15) that there exist $\xi_{n,1} \in (0,1), \xi_{n,2} \in (0,1)$ such that

$$\begin{aligned} \underline{c} d_n^2 &\le E(U_n) - E(U^*) \\ &= \nabla^2_G E\big(\exp_{[U^*]}(\xi_{n,1} D_n)\big)\left[\tau_{\xi_{n,1}} D_n, \tau_{\xi_{n,1}} D_n\right] \\ &\le \bar{c} d_n^2, \end{aligned} \tag{3.16}$$

and

$$\begin{aligned} \|\nabla_G E(U_n)\| &= \left\|\tau^{-1}_{\xi_{n,2}} \nabla^2_G E\big(\exp_{[U^*]}(\xi_{n,2} D_n)\big)\left[\tau_{\xi_{n,2}} D_n\right]\right\| \\ &= \left\|\nabla^2_G E\big(\exp_{[U^*]}(\xi_{n,2} D_n)\big)\left[\tau_{\xi_{n,2}} D_n\right]\right\| \ge \underline{c} d_n. \end{aligned} \tag{3.17}$$

Combining (3.16) and (3.17), we see that

$$\|\nabla_G E(U_n)\|^2 \ge \frac{\underline{c}^2}{\bar{c}}\big(E(U_n) - E(U^*)\big). \tag{3.18}$$

Hence, we obtain from the fact that $\{\Delta t_n\}_{n=0}^\infty$ is bounded from below that

$$\begin{aligned} E(U_{n+1}) - E(U^*) &\le -\eta \Delta t_n \|\nabla_G E(U_n)\|^2 + E(U_n) - E(U^*) \\ &\le \left(1 - \eta\tau\frac{\underline{c}^2}{\bar{c}}\right)\big(E(U_n) - E(U^*)\big). \end{aligned}$$

Finally, we complete the proof by using (3.16) and choosing

$$\nu = 1 - \eta\tau\frac{\underline{c}^2}{\bar{c}}, \quad C_1 = \bar{c}, \quad C_2 = \left(\frac{\bar{c}}{\underline{c}}\right)^{\frac{1}{2}}.$$

The proof is complete. □

4. Some typical orthogonality preserving schemes

In the previous section, we propose and analyze a general framework for interpolation based orthogonality preserving schemes for discretizing the gradient flow based Kohn-Sham DFT model (2.9). In that framework, how to determine the specific form of the auxiliary mapping U^{Aux} and the time step size Δt_n are not given. The specific form and the efficiency of Algorithm 1 depend strongly on the definition of the auxiliary mapping U^{Aux} and the choice of the time step size. For example, if the time step size Δt_n is chosen to be too large, Assumption 3.1 may not hold, which may lead to divergence. On the contrary, Theorem 3.3 indicates that tiny step sizes will lead to slow convergence. In this section, we will provide some choices for the auxiliary mapping U^{Aux}, and propose an adaptive approach for determining the time step sizes. We will prove that our approach can produce time step sizes which can not only satisfy Assumption 3.1 but also avoid slow convergence.

4.1. Auxiliary mapping U^{Aux}

We see from the previous discussion that Algorithm 1 gives a general framework of orthogonality preserving numerical schemes for solving (3.1), which provides at least a subsequence that converges to the equilibrium point under some mild assumptions. All our analysis in Section 3 is independent of the specific form of U^{Aux} at each interval $[t_n, t_{n+1})$. However, the choice of the auxiliary mapping is one of the keys when we carry out Algorithm 1. Here, we provide some potential choices for the auxiliary mapping $U^{\text{Aux}}(t)$.

Choice 1. $U^{\text{Aux}}(t) = (1-\alpha_n)U_n + \alpha_n \tilde{U}(t)$, $\alpha_n \in [0,1]$, $t \in [t_n, t_{n+1})$.

If we use Crank-Nicolson's strategy [9], which is a widely used second order scheme in time, to discretize (2.9), then we have

$$U_{n+1} = \left(I_N + \frac{\Delta t}{2}\mathcal{A}_{U_{n+1}}\right)^{-1}\left(I_N - \frac{\Delta t}{2}\mathcal{A}_{U_n}\right)U_n. \tag{4.1}$$

However, it may not preserve the orthogonality of orbitals. Notice that if we choose $\alpha_n = 0$ in Choice 1, then the orbitals can be updated by

$$U_{n+1} = \left(I_N + \frac{\Delta t}{2}\mathcal{A}_{U_n}\right)^{-1}\left(I_N - \frac{\Delta t}{2}\mathcal{A}_{U_n}\right)U_n, \tag{4.2}$$

which preserves the orthogonality automatically and is an approximation of Crack-Nicolson scheme (4.1) simply by substituting $\mathcal{A}_{U_{n+1}}$ with $\mathcal{A}_{U_n}$ in (4.1). Hence, we may denote the auxiliary mapping in this case as $U^{\text{Aux}}_{\text{CN}}(t) = U_n$. Besides, if α_n is chosen to be $1/2$, then we have

$$U^{\text{Aux}}(t) = \frac{(\tilde{U}(t) + U_n)}{2}.$$

We can see that in this case, the updating formula is the same as the midpoint scheme studied in [12]. Hence, we denote

$$U_{\text{Mid}}^{\text{Aux}}(t)=\frac{\tilde{U}(t)+U_n}{2}.$$

Therefore, the framework that we proposed (Algorithm 1) contains both the Crank-Nicolson like scheme (4.2) and the midpoint scheme [12].

Under the classification mentioned in this paper, we see that the midpoint scheme is implicit and can not be easily carried out. Instead, Dai *et al.* also proposed an explicit approximation to the midpoint scheme based on the Picard iteration [12]. More precisely, the midpoint scheme can be replaced approximately by the iterative formulae

$$U_{n+1/2}^{m}(t)=\left(I+\frac{t-t_n}{2}\mathcal{A}_{U_{n+1/2}^{m-1}(\Delta t_n)}\right)^{-1}U_n,\quad m=1,2,\dots,$$

where $U_{n+1/2}^{0}=U_n$. This gives us the following choice.

Choice 2. $U^{\text{Aux}}(t)=U_{n+1/2}^{m}(t)=:U_{\text{aMid-m}}^{\text{Aux}}(t),t\in[t_n,t_{n+1}),\forall m\in\mathbb{N}_0$.

It is easy to check that $U_{\text{aMid-m}}^{\text{Aux}}(t_n)=U_n$. Replacing the midpoint $(U_n+\tilde{U}(t))/2$ by the approximated midpoints $U_{\text{aMid-m}}^{\text{Aux}}(t)$ in the midpoint scheme, we obtain a set of explicit schemes and name them as approximated midpoint schemes. They are also included in our interpolation based schemes.

We can of course construct some simpler explicit schemes, e.g., the following one.

Choice 3. Let

$$U^{\text{Aux}}(t)=U_n-m_n(t-t_n)\nabla_G E(U_n),\quad t\in[t_n,t_{n+1})\tag{4.3}$$

or

$$U^{\text{Aux}}(t)=2\big(I+m_n(t-t_n)\mathcal{A}_{U_n}\big)^{-1}U_n-U_n,\quad t\in[t_n,t_{n+1}),\tag{4.4}$$

where m_n can be arbitrary real number.

There are also many other explicit schemes and we will not go further into them. With regard to implicit schemes, we propose the following example motivated by the Verlet algorithm [40]. Consider the first order Taylor expansion of $U(t)$ at $t_{n+1/2}=(t_n+t_{n+1})/2$, that is,

$$\begin{aligned}U_{n+1}&\approx U\left(t_{n+1/2}+\frac{\Delta t_n}{2}\right)\approx U(t_{n+1/2})+\frac{\Delta t_n}{2}\dot{U}(t_{n+1/2}),\\ U_n&\approx U\left(t_{n+1/2}-\frac{\Delta t_n}{2}\right)\approx U(t_{n+1/2})-\frac{\Delta t_n}{2}\dot{U}(t_{n+1/2}).\end{aligned}$$

It can be observed that the midpoint scheme uses $(U_n+U_{n+1})/2$ to approximate $U_{n+1/2}$ with linear accuracy. We may further consider the second order Taylor expansion of $U(t)$, which is formulated as

$$\begin{aligned}
U_{n+1} &\approx U\left(t_{n+1/2}+\frac{\Delta t_n}{2}\right) \approx U(t_{n+1/2})+\frac{\Delta t_n}{2}\dot{U}(t_{n+1/2})+\frac{\Delta t_n^2}{8}\ddot{U}(t_{n+1/2}),\\
U_n &\approx U\left(t_{n+1/2}-\frac{\Delta t_n}{2}\right) \approx U(t_{n+1/2})-\frac{\Delta t_n}{2}\dot{U}(t_{n+1/2})+\frac{\Delta t_n^2}{8}\ddot{U}(t_{n+1/2}),
\end{aligned}$$

where

$$\begin{aligned}
\ddot{U}(t) &= -\frac{d}{dt}\nabla_G E\big(U(t)\big)\\
&= \big(I-U(t)U(t)^T\big)\,\nabla^2 E\big(U(t)\big)\,\big[\nabla_G E\big(U(t)\big)\big]\\
&\quad+\Big(\nabla_G E\big(U(t)\big)U(t)^T+U(t)\nabla_G E\big(U(t)\big)^T\Big)\,\nabla E\big(U(t)\big)\\
&=: G\big(U(t)\big),
\end{aligned} \tag{4.5}$$

based on which a higher order approximation of the midpoint can be obtained.

Choice 4.

$$U^{\text{Aux}}(t)=\frac{\tilde{U}(t)+U_n}{2}-\frac{(t-t_n)^2}{8}G\left(\frac{\tilde{U}(t)+U_n}{2}\right)=: U^{\text{Aux}}_{\text{Verlet}}(t), \quad t\in[t_n,t_{n+1}).$$

Here, the operator G can be defined as (4.5) or it can be chosen as some approximations of (4.5).

We should emphasize that there are many different choices for the auxiliary mapping $U^{\text{Aux}}(t)$. Each of them will result in a specific orthogonality preserving scheme for the discretized Kohn-Sham model. The difference lies in the efficiency, which will be further studied in our future work.

4.2. Time step sizes

The choice of the time step sizes is of great importance in the discretization of time dependent problems, on which many studies have been done in literature (see, e.g., [23, 35]). In the numerical analysis for Algorithm 1 provided in Section 3, we require that the time step sizes satisfy Assumption 3.1. Here, we provide a concrete method to help us judge whether or not a preset step size Δt_n satisfies the energy decrease property of the gradient flow based model (2.11), whose key idea is to use the second-order Taylor expansion to approximate the energy functional $E(\tilde{U}(t_n+\Delta t))$. Note that a similar idea has been used in [13].

By using the second-order Taylor expansion, we have the following approximation:

$$\begin{aligned}
E\big(\tilde{U}(t_n+\Delta t)\big) \approx{}& E\big(\tilde{U}(t_n)\big)+\Delta t\big\langle\nabla_G\big(E\big(\tilde{U}(t_n)\big)\big),\tilde{U}'(t_n)\big\rangle\\
&+\frac{\Delta t^2}{2}\nabla_G^2 E\big(\tilde{U}(t_n)\big)\big[\tilde{U}'(t_n),\tilde{U}'(t_n)\big].
\end{aligned} \tag{4.6}$$

From the definition of $\tilde{U}$, we see that

$$\tilde{U}(t_n) = U_n, \quad \tilde{U}'(t_n) = -\nabla_G E(U_n),$$

and thus (4.6) becomes

$$\begin{aligned} E\big(\tilde{U}(t_n + \Delta t)\big) \approx\ & E(U_n) - \Delta t \|\nabla_G (E(U_n)\|^2 \\ & + \frac{\Delta t^2}{2} \nabla_G^2 E(U_n)\big[\nabla_G\big(E(U_n)\big), \nabla_G(E(U_n)\big]. \end{aligned} \tag{4.7}$$

Inserting (4.7) into (3.6) in Assumption 3.1, we have the following inequality for the time step size Δt:

$$\frac{\|\nabla_G E(U_n)\|^2 - \Delta t/2 \nabla_G^2 E(U_n)\big[\nabla_G E(U_n), \nabla_G E(U_n)\big]}{\|\nabla_G E(U_n)\|^2} \geq \eta.$$

Therefore, for a given Δt at the n-th iteration, we define the following indicator:

$$\zeta_n(\Delta t) = \frac{\|\nabla_G E(U_n)\|^2 - \Delta t/2 \nabla_G^2 E(U_n)\big[\nabla_G E(U_n), \nabla_G E(U_n)\big]}{\|\nabla_G E(U_n)\|^2} \tag{4.8}$$

to tell us if it is a good step size. If $\zeta_n(\Delta t) \geq \eta$, we consider Δt as a satisfactory time step and accept it. Otherwise, we instead choose Δt_n to be the approximated minimizer of $E(\tilde{U}(t_n + \Delta t))$ with respect to Δt at the n-th iteration. That is, we choose

$$\Delta t_n = \min\left\{ \frac{\|\nabla_G E(U_n)\|^2}{\nabla_G^2 E(U_n)[\nabla_G E(U_n), \nabla_G E(U_n)]}, \frac{\theta_n}{\|\nabla_G E(U_n)\|} \right\},$$

which is the minimizer of the right-hand side of (4.7) in a small neighbourhood of 0, to be our final step size.

In summary, we obtain an adaptive strategy to get the time step sizes which will be proved to satisfy Assumption 3.1. With this strategy, the corresponding interpolation based scheme reads as the following Algorithm 2, where $\delta t_{\min}$ and $\delta t_{\max}$ are the preset bound for the initial step sizes.

For Algorithm 2, we have the following theorem, which shows the convergence of our interpolation based scheme with adaptive step sizes.

Theorem 4.1. *If Assumption* 3.2 *holds and the initial guess* $[U_0] \in B([U^*], \delta_1) \subset \mathcal{G}_{N_g}^N$, *then there exists* $\{\theta_n\}_{n\in\mathbb{N}_0}$ *such that for the sequence* $\{U_n\}_{n\in\mathbb{N}_0}$ *generated by Algorithm* 2, *there holds either* $\nabla_G E(U_n) = 0$ *for some* $n \in \mathbb{N}_0$ *or*

$$\liminf_{n\to\infty} \|\nabla_G E(U_n)\| = 0. \tag{4.9}$$

Furthermore, there also holds that

$$\liminf_{n\to\infty} \operatorname{dist}_{geo}(U_n, U^*) = 0.$$

Algorithm 2: Interpolation based scheme with adaptive step sizes.

```
Given $\epsilon, \delta t_{\min}, \delta t_{\max} > 0$, $\eta \in (0, 1/2)$, initial data $U_0 \in \mathcal{M}^N_{N_g}$, calculate the
  gradient $\nabla_G E(U_0)$ and set $n = 0$, $t_0 = 0$.
while $\|\nabla_G E(U_n)\| > \epsilon$ do
    Choose $\theta_n \in (0, 1)$ and the initial guess $\Delta t_n^{initial} \in [\delta t_{\min}, \delta t_{\max}]$ by some
      specific strategy.
    Let $\Delta t_n = \Delta t_n^{initial}$.
    if $\zeta_n(\Delta t_n) < \eta$ or $\Delta t_n > \theta_n / \|\nabla_G E(U_n)\|$ then
        $\Delta t_n = \min\left\{ \frac{\|\nabla_G E(U_n)\|^2}{\nabla^2_G E(U_n)[\nabla_G E(U_n), \nabla_G E(U_n)]}, \frac{\theta_n}{\|\nabla_G E(U_n)\|} \right\}$.
    Set $t_{n+1} = t_n + \Delta t_n$.
    Define $U^{\mathrm{Aux}}(t)$ on the interval $[t_n, t_{n+1})$ such that $U^{\mathrm{Aux}}(t_n) = U_n$.
    Update $U_{n+1} = \lim_{t \to t_{n+1}^-} \tilde{U}(t)$ with $\tilde{U}(t)$ satisfying (3.2).
    Let $n = n + 1$, calculate the gradient $\nabla_G E(U_n)$.
```

Proof. To simplify the notation, we denote $D_n = -\nabla_G E(U_n)$. We see that the time step size Δt_n given by Algorithm 2 should satisfy $\Delta t_n \|D_n\| \le \theta_n$ and

$$\Delta t_n \|D_n\|^2 + \frac{\Delta t_n^2}{2} \nabla^2_G E(U_n)[D_n, D_n] \le \eta \Delta t_n \|D_n\|^2, \quad \forall n \in \mathbb{N}_0.$$

Define

$$\theta_n = \sup \left\{ \tilde{\theta}_n : E\big(\tilde{U}(t_n + \Delta t)\big) - E(U_n) - \Delta t \|D_n\|^2 - \frac{\Delta t^2}{2} \nabla^2_G E(U_n)[D_n, D_n] \right.$$
$$\left. \le -\frac{\eta \Delta t \|D_n\|^2}{2}, \ \forall \Delta t \le \frac{\tilde{\theta}_n}{\|D_n\|} \right\} \ge 0.$$

Then, we obtain from the definition of $E(U_{n+1})$ and θ_n that

$$E(U_{n+1}) - E(U_n) \le \frac{\eta}{2} \Delta t_n \|D_n\|^2, \quad \forall n \in \mathbb{N}_0,$$

i.e., (3.6) holds.

As for the condition (3.7), we see that Δt_n has only three possible values, that is,

$$\Delta t_n = \max\big(t_n^{\text{initial}}, \delta t_{\min}\big), \quad \Delta t_n = \frac{\|D_n\|^2}{\nabla^2_G E(U_n)[D_n, D_n]},$$

or

$$\Delta t_n = \frac{\theta_n}{\|D_n\|}.$$

So there is at least one infinite subsequence of $\{n_j\}_{j=0}^{\infty}$, which is, without loss of generality, also denoted by $\{n\}_{n=0}^{\infty}$ such that

Case 1. $\Delta t_n = \max\left(\Delta t_n^{\text{initial}}, \delta t_{\min}\right)$. We have immediately that

$$\sum_{n=0}^{\infty} \Delta t_n \geq \sum_{j=0}^{\infty} \Delta \delta t_{\min} = +\infty.$$

Case 2. $\Delta t_n = \|D_n\|^2/(\nabla_G^2 E(U_n)[D_n, D_n])$. We obtain from Assumption 3.2 that $\Delta t_n \geq 1/\bar{c}$ and hence

$$\sum_{n=0}^{\infty} \Delta t_n = +\infty.$$

Case 3. $\Delta t_n = \theta_n/\|D_n\|$. If $\liminf_{n\to\infty} \Delta t_n > 0,$ then (3.7) is satisfied. Otherwise, there exists a subsequence of $\{\Delta t_n\}_{n\in\mathbb{N}_0}$, which is also denoted by $\{\Delta t_n\}_{n\in\mathbb{N}_0}$ such that $\lim_{n\to\infty} \Delta t_n = 0$. This simply leads to $\lim_{n\to\infty} \theta_n = 0$ since $\|D_n\|$ is bounded from above.

We have that for all $n \in \mathbb{N}_0$, there hold

$$\begin{aligned}
&E\big(\tilde{U}(t_n + \Delta t)\big) - E(U_n) - \Delta t\|D_n\|^2 - \frac{\Delta t^2}{2}\nabla_G^2 E(U_n)[D_n, D_n] \\
&= E\big(\tilde{U}(t_n + \Delta t)\big) - E\big(\exp_{[U_n]}(\Delta t D_n)\big) + E\big(\exp_{[U_n]}(\Delta t D_n)\big) - E(U_n) \\
&\quad + \Delta t\|D_n\|^2 - \frac{\Delta t^2}{2}\nabla_G^2 E(U_n)[D_n, D_n] =: T_n^{(1)} + T_n^{(2)},
\end{aligned}$$

where

$$\begin{aligned}
T_n^{(1)} &= E\big(\tilde{U}(t_n + \Delta t)\big) - E\big(\exp_{[U_n]}(\Delta t D_n)\big), \\
T_n^{(2)} &= E\big(\exp_{[U_n]}(\Delta t D_n)\big) - E(U_n) + \Delta t\|D_n\|^2 - \frac{\Delta t^2}{2}\nabla_G^2 E(U_n)[D_n, D_n].
\end{aligned}$$

We see from Remark 3.2 that there exists a geodesic $[\exp_{[\tilde{U}(t_n+\Delta t)]}(t\hat{D})]$ such that

$$\begin{aligned}
&\exp_{[\tilde{U}(t_n+\Delta t)]}(0) = \tilde{U}(t_n + \Delta t), \\
&\left[\exp_{[\tilde{U}(t_n+\Delta t)]}(\hat{D})\right] = \left[\exp_{[U_n]}(\Delta t D_n)\right],
\end{aligned}$$

and obtain from (3.15) that

$$\begin{aligned}
\big|T_n^{(1)}\big| &= \big|E\big(\exp_{[\tilde{U}(t_n+\Delta t)]}(0\hat{D})\big) - E\big(\exp_{[\tilde{U}(t_n+\Delta t)]}(\hat{D})\big)\big| \\
&= \big|\big\langle \nabla_G E\big(\exp_{[\tilde{U}(t_n+\Delta t)]}(\xi\hat{D})\big), \tau_{(\tilde{U}(t_n+\Delta t),\hat{D},\xi)}\hat{D}\big\rangle\big| \\
&\leq \big\|\nabla_G E\big(\exp_{[\tilde{U}(t_n+\Delta t)]}(\xi\hat{D})\big)\big\|\big\|\tau_{(\tilde{U}(t_n+\Delta t),\hat{D},\xi)}\hat{D}\big\| \leq C\|\hat{D}\|,
\end{aligned}$$

where Assumption 3.2 and (3.12) are used in the last inequality.

From (3.14), we have that

$$\begin{aligned}\|\hat{D}\| &= \mathrm{dist}_{geo}\left(\left[\tilde{U}(t_n+\Delta t)\right], \left[\exp_{[U_n]}(\Delta t D_n)\right]\right)\\ &\le 2\,\mathrm{dist}_{cF}\left(\left[\tilde{U}(t_n+\Delta t)\right], \left[\exp_{[U_n]}(\Delta t D_n)\right]\right)\\ &\le 2\big\|\tilde{U}(t_n+\Delta t) - \exp_{[U_n]}(\Delta t D_n)\big\|\\ &\le 2\big(\|\tilde{U}(t_n+\Delta t) - U_n - \Delta t D_n\| + \|\exp_{[U_n]}(\Delta t D_n) - U_n - \Delta t D_n\|\big).\end{aligned}$$

Notice that $\tilde{U}(t)$ and $\exp_{[U_n]}(tD_n)$ satisfy

$$\tilde{U}(t_n) = U_n, \quad \tilde{U}'(t_n) = D_n$$

and

$$\exp_{[U_n]}(0) = U_n, \quad \exp_{[U_n]}{}'(0) = D_n,$$

we have $\|\hat{D}\| = o(\Delta t)$. If the sequence $\{\|D_n\|\}_{n\in\mathbb{N}_0}$ is not bounded from below, then we complete the proof. Otherwise, we obtain

$$T_n^{(1)} = o\big(\Delta t\|D_n\|\big) \tag{4.10}$$

since $\|D_n\|$ is bounded from both above and below. As for $T_n^{(2)}$, the relation (3.15) gives that

$$T_n^{(2)} = o\big(\Delta t^2\|D_n\|^2\big). \tag{4.11}$$

Combining (4.10) and (4.11), we arrive at

$$\begin{aligned}&E\big(\tilde{U}(t_n+\Delta t)\big) - E(U_n) - \Delta t\|D_n\|^2 - \frac{\Delta t^2}{2}\nabla_G^2 E(U_n)[D_n,D_n]\\ &= T_n^{(1)} + T_n^{(2)} = o\big(\Delta t\|D_n\|\big), \quad \forall n\in\mathbb{N}_0.\end{aligned}$$

Note that the definition of θ_n implies that for all n, there exists

$$\Delta t_n^* \in \left(\frac{\theta_n}{\|D_n\|}, \frac{\theta_n + 1/n}{\|D_n\|}\right)$$

such that

$$\begin{aligned}o\big(\Delta t_n^*\|D_n\|\big) &= E\big(\tilde{U}(t_n+\Delta t_n^*)\big) - E(U_n) + \Delta t_n^*\|D_n\|^2\\ &\quad - \frac{\Delta t_n^{*\,2}}{2}\nabla_G^2 E(U_n)[D_n,D_n]\\ &> \frac{\eta\Delta t_n^*\|D_n\|^2}{2}.\end{aligned} \tag{4.12}$$

Hence, it is easy to see that

$$0 \le \lim_{n\to\infty} t_n^*\|D_n\| \le \lim_{n\to\infty}\left(\theta_n + \frac{1}{n}\right) = 0.$$

Finally, by letting $n\to\infty$ in (4.12) we obtain that

$$0 \ge \lim_{n\to\infty}\frac{\eta}{2}\|D_n\|,$$

which together with (3.17) completes the proof. □

5. Numerical experiments

In this section, we apply one of our proposed schemes to solve the discretized Kohn-Sham DFT model for some typical systems, including benzene (C_6H_6), aspirin $(C_9H_8O_5)$ and Fullerin (C_{60}), to validate our theoretical results. More specifically, we test the scheme (Algorithm 2) with auxiliary mapping (4.3) and with m_n being chosen as $1/2$. All of our experiments are carried out on LSSC-IV cluster and the coding is built based on the software package Octopus* (Version 4.0.1). Among all our experiments, we set $\eta = 1\text{e}{-4}$, $\epsilon = 1\text{e}{-12}$, $\delta t_{\min} = 1\text{e}{-20}$, and $\Delta t_n^{initial} = 0.1$, $\theta_n = 0.8$, for all n. Here and hereafter, we denote this specific scheme as GF-EX scheme.

We first test the orthogonality preserving property of our scheme. To this end, we define the orthogonality violation of the iterative orbital U_n at the n-th iteration as

$$\varepsilon_n = \left\| U_n^T U_n - I_N \right\|_F$$

and show the curves for $\{\varepsilon_n\}_n$ in Fig. 1, of which the x-axis stands for the number of iteration n and the y-axis is the value of ε_n.

It can be observed from Fig. 1 that the orthogonality violations for all tested systems always lie in the interval (1e-15,1e-13) during the iteration, which indicates that the GF-EX scheme indeed preserves the orthogonality of iterative orbitals well.

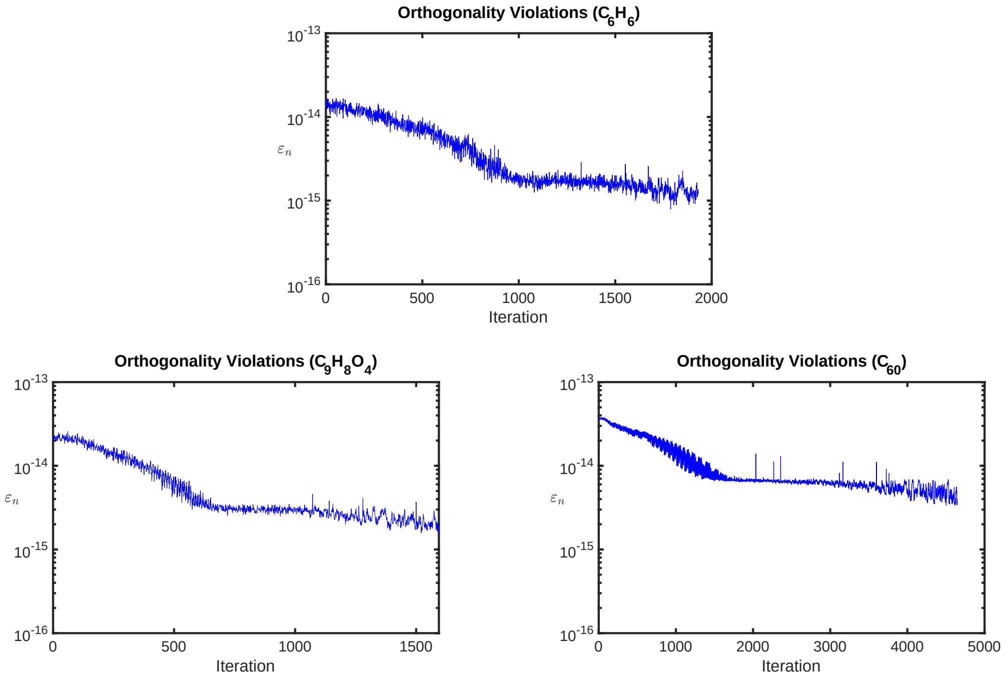

Figure 1: Orthogonality violations obtained by GF-EX for different systems.

*Octopus: `octopus-code.org/wiki/Main_Page`

Then we show the convergent states obtained by our scheme in Table 1, in which the reference ground state energy $E_{\min}$ is obtained by the SCF iterations provided by Octopus and U_{final} stands for the orbitals obtained at the last iteration when the iteration meets the stopping criterion.

We see from Table 1 that our scheme can indeed produce approximations that converge to the ground state. Figs. 2-3 illustrate the convergence curves for the error of energy and the norm of $\nabla_G E$ obtained by our scheme, respectively, which give an intuitive look for the numerical behaviour of the GF-EX scheme.

In the energy plots (Fig. 2), the y-axis indicates the energy difference between $E(U_n)$ and $E_{\min}$. Among both Figs. 2 and 3, the red lines show the asymptotical infimum of iterations, which are defined as

$$\begin{aligned}
&\inf_n \{E(U_n) - E_{\min}\} = \min_{i\in\{1,2,\ldots,n\}} \{E(U_i) - E_{\min}\},\\
&\inf_n \|\nabla_G E(U_n)\| = \min_{i\in\{1,2,\ldots,n\}} \|\nabla_G E(U_i)\|.
\end{aligned}$$

We observe that the red line in Fig. 2 terminates earlier than the blue one. The reason is that we have achieved a lower energy than the reference energy $E_{\min}$ during the iteration, which makes $\inf_n\{E(U_n) - E_{\min}\}$ negative and thus can no longer be shown in the log scale plots. The above two figures show the convergence of both the energy and the norm of gradient clearly, which is consistent to our theory.

Table 1: Numerical results obtained by the scheme GF-EX.

System	Reference energy $E_{\min}$ (a.u.)	$E(U_{\text{final}})$ (a.u.)	$\|\nabla_G E(U_{\text{final}})\|$
Benzene (C_6H_6)	-3.74246025E+01	-3.74246025E+01	9.92E-13
Aspirin ($C_9H_8O_4$)	-1.20214764E+02	-1.20214764E+02	6.69E-13
Fullerin (C_{60})	-3.42875137E+02	-3.42875137E+02	9.91E-13

6. Concluding remarks

In this paper, we have proposed and analyzed a general framework of orthogonality preserving schemes for approximating the Kohn-Sham orbitals, from which we can obtain a class of orthogonality preserving schemes. We have proved the convergence and derived the local exponential convergence rate of the framework under some mild and reasonable assumptions. In addition, we have provided some typical choices for the auxiliary mapping which lead to several orthogonality preserving schemes. We have also presented an efficient approach to obtain the desired time step sizes that satisfy the assumptions required in our analysis. We have applied one of the explicit schemes that we proposed as an example to verify our theory. Due to the great flexibility on choosing both auxiliary mapping and step sizes in our framework, we will systematically study, apply and compare the schemes generated by our framework based on numerical experiments on electronic structure calculations in our future work.

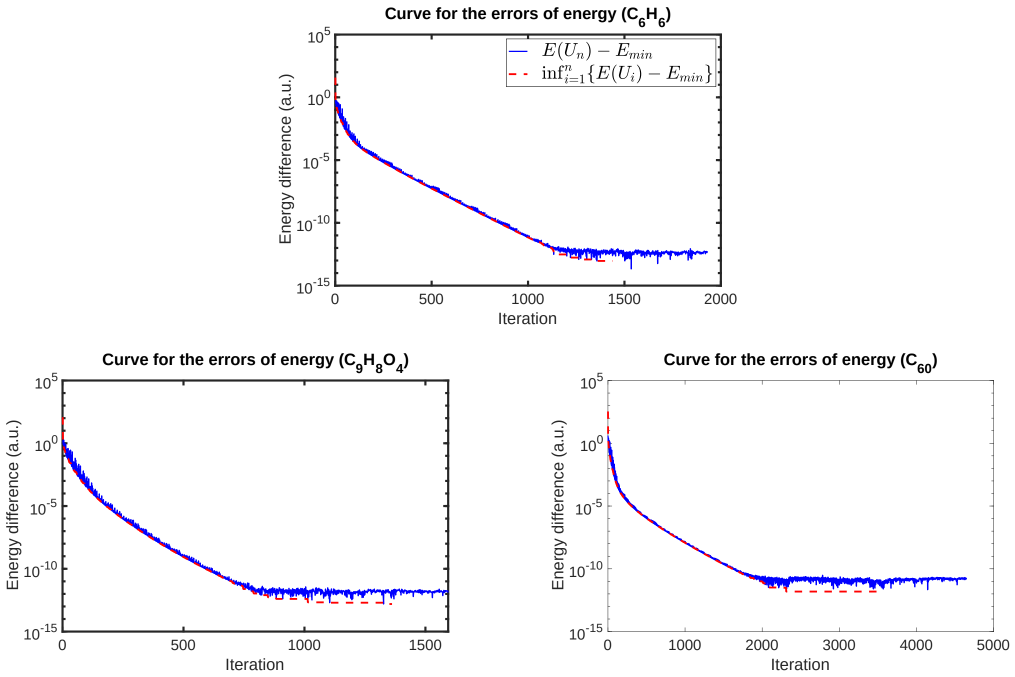

Figure 2: Curves for the error of energy obtained by GF-EX for different systems.

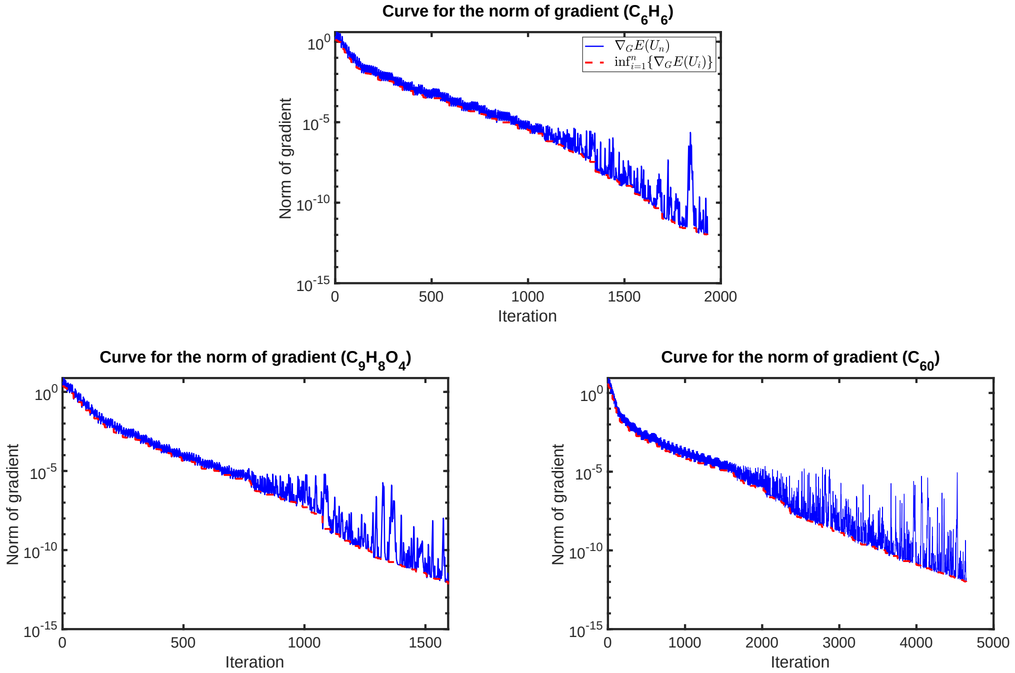

Figure 3: Curves for $\|\nabla_G E\|$ obtained by GF-EX for different systems.

Acknowledgments

The authors would like to thank the anonymous referees for their helpful comments and suggestions that enrich the content and improve the presentation of the manuscript.

This work was supported by the National Key R & D Program of China under grants 2019YFA0709600, 2019YFA0709601, and the National Natural Science Foundation of China under grant 12021001.

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, 2008.

[2] D. G. Anderson, *Iterative procedures for nonlinear integral equations*, J. ACM 12 (1965), 547–560.

[3] Z. Bai, R.-C. Li, and D. Lu, *Optimal convergence rate of self-consistent field iteration for solving eigenvector-dependent nonlinear eigenvalue problems*, arXiv:2009.09022 (2020).

[4] E. Cances, *Self-consistent field algorithms for Kohn-Sham models with fractional occupation numbers*, J. Chem. Phys. 114 (2001), 10616–10622.

[5] E. Cances, R. Chakir, and Y. Maday, *Numerical analysis of the planewave discretization of some orbital-free and Kohn-Sham models*, ESAIM Math. Model. Numer. Anal. 46 (2012), 341–388.

[6] E. Cances, G. Kemlin, and A. Levitt, *Convergence analysis of direct minimization and self-consistent iterations*, SIAM J. Matrix Anal. Appl. 42 (2021), 243–274.

[7] H. Chen, X. Dai, X. Gong, L. He, and A. Zhou, *Adaptive finite element approximations for Kohn- Sham models*, Multiscale Model. Simul. 12 (2014), 1828–1869.

[8] H. Chen, X. Gong, L. He, Z. Yang, and A. Zhou, *Numerical analysis of finite dimensional approximations of Kohn-Sham equations*, Adv. Comput. Math. 38 (2013), 225–256.

[9] J. Crank and P. Nicolson, *A practical method for numerical evaluation of solutions of partial differential equations of the heat conduction type*, Math. Proc. Cambridge Philos. Soc. 43 (1947), 50–67.

[10] X. Dai, Z. Liu, L. Zhang, and A. Zhou, *A conjugate gradient method for electronic structure calculations*, SIAM J. Sci. Comput. 39 (2017), 2702–2740.

[11] X. Dai, Z. Liu, X. Zhang, and A. Zhou, *A parallel orbital-updating based optimization method for electronic structure calculations*, J. Comput. Phys. 445 (2021), 110622.

[12] X. Dai, Q. Wang, and A. Zhou, *Gradient flow based discretized Kohn-Sham density functional theory*, Multiscale Model. Simul. 18 (2020), 1621–1663.

[13] X. Dai, L. Zhang, and A. Zhou, *An adaptive step size strategy for orthogonality constrained line search methods*, arXiv:1906.02883, (2019).

[14] X. Dai and A. Zhou, *Finite element methods for electronic structure calculations* (in Chinese), SCIENTIA SINICA Chimica 45 (2015), 800–811.

[15] J. Ding and A. Zhou, *A spectrum theorem for perturbed bounded linear operators*, Appl. Math. Comput. 201 (2008), 723–728.

[16] A. Edelman, T. A. Arias, and S. T. Smith, *The geometry of algorithms with orthogonality constraints*. SIAM J. Matrix Anal. Appl. 20 (1998), 303–353.

[17] C. Freysoldt, S. Boeck, and J. Neugebauer, *Direct minimization technique for metals in density functional theory*, Phys. Rev. B 79 (2009), 241103.

[18] B. GAO, X. LIU, X. CHEN, AND Y. YUAN, *A new first-order framework for orthogonal constrained optimization problems*, SIAM J. Optim. 28 (2017), 302–332.

[19] B. GAO, X. LIU, AND Y. YUAN, *Parallelizable algorithms for optimization problems with orthogonality constraints*, SIAM J. Sci. Comput. 41 (2019), A1949–A1983.

[20] P. HOHENBERG AND W. KOHN, *Inhomogeneous electron gas*, Phys. Rev. B. 136 (1964), 864–871.

[21] D. D. JOHNSON, *Modified Broyden's method for accelerating convergence in self-consistent calculations*, Phys. Rev. B 38 (1988), 12807–12813.

[22] W. KOHN AND L. J. SHAM, *Self-consistent equations including exchange and correlation effects*, Phys. Rev. A. 140 (1965), 4743–4754.

[23] H. LIAO, T. TANG, AND T. ZHOU, *A second-order and nonuniform time-stepping maximum-principle preserving scheme for time-fractional Allen-Cahn equations*, J. Comput. Phys. 414 (2020), 109473.

[24] L. LIN, J. LU, AND L. YING, *Numerical methods for Kohn-Sham density functional theory*, Acta Numer. 28 (2019), 405–539.

[25] 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), S277–S298.

[26] X. LIU, X. WANG, Z. WEN, AND Y. YUAN, *On the convergence of the self-consistent field iteration in Kohn-Sham density functional theory*, SIAM J. Matrix Anal. Appl. 35 (2014), 546–558.

[27] X. LIU, Z. WEN, X. WANG, M. ULBRICH, AND Y. YUAN, *On the analysis of the discretized Kohn-Sham density functional theory*, SIAM J. Numer. Anal. 53 (2015), 1758–1785.

[28] M. A. L. MARQUES, N. T. MAITRA, F. M. S. MOGUEIRA, E. K. U. GROSS, AND A. RUBIO, Eds., *Fundamentals of Time-Dependent Density Functional Theory*, Lecture Notes in Physics, Vol. 837, Springer, 2012.

[29] R. MARTIN, *Electronic Structure: Basic Theory and Practical Methods*, Cambridge University Press, 2004.

[30] 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), 1337.

[31] R. G. PARR AND W. YANG, *Density-Functional Theory of Atoms and Molecules*, Oxford University Press, 1994.

[32] M. C. PAYNE, M. P. TETER, D. C. ALLEN, T. A. ARIAS, AND J. D. JOANNOPOULO, *Iterative minimization techniques for ab initio total energy calculation: Molecular dynamics and conjugate gradients*, Rev. Mod. Phys. 64 (1992), 1045–1097.

[33] P. PULAY, *Convergence acceleration of iterative sequences: The case of SCF iteration*, Chem. Phys. Lett. 73 (1980), 393–398.

[34] P. PULAY, *Improved SCF convergence acceleration*, J. Comput. Chem. 3 (1982), 556–560.

[35] Z. QIAO, Z. ZHANG, AND T. TANG, *An adaptive time-stepping strategy for the molecular beam epitaxy models*, SIAM J. Sci. Comput. 33 (2011), 1395–1414.

[36] E. RUNGE AND E. K. U. GROSS, *Density functional theory for time-dependent systems*, Phys. Rev. Lett. 52 (1984), 997–1000.

[37] Y. SAAD, J. R. CHELIKOWSHY, AND S. M. SHONTZ, *Numerical methods for electronic structure calculations of materials*, SIAM Rev. 52 (2010), 3–54.

[38] R. SCHNEIDER, T. ROHWEDDER, A. NEELOV, AND J. BLAUERT, *Direct minimization for calculating invariant subspaces in density fuctional computations of the electronic structure*, J. Comput. Math. 27 (2009), 360–387.

[39] S. T. SMITH, *Optimization techniques on Riemannian manifolds*, in: *Fields Institute Com-*

munications, Vol. 3, AMS, (1994), 113–146.

[40] L. VERLET, *Computer 'experiments' on classical fluids. 1. Thermodynamical properties of LennardJones molecules*, Phys. Rev. 159 (1967), 98–103.

[41] C. YANG, W. GAO, AND J. MEZA, *On the convergence of the self-consistent field iteration for a class of nonlinear eigenvalue problems*, SIAM J. Matrix Anal. Appl. 30 (2009), 1773–1788.

[42] C. YANG, J. C. MEZA, AND L. WANG, *A trust region direct constrained minimization algorithm for the Kohn-Sham equation*, SIAM J. Sci. Comput. 29 (2007), 1854–1875.

[43] L. YANG, Y. SHEN, Z. HU, AND G. HU, *An implicit solver for the time-dependent Kohn-Sham equation*, Numer. Math. Theor. Meth. Appl. 14 (2020), 261–284.

[44] D. ZHANG, L. SHEN, A. ZHOU, AND X. GONG, *Finite element method for solving Kohn-Sham equations based on self-adaptive terahedral mesh*, Phys. Lett. A 372 (2008), 5071–5076.

[45] X. ZHANG, J. ZHU, Z. WEN, AND A. ZHOU, *Gradient type optimization methods for electronic structure calculations*, SIAM J. Sci. Comput. 36 (2014), 265–289.

[46] Z. ZHAO, Z. BAI, AND X. JIN, *A Riemannian Newton algorithm for nonlinear eigenvalue problems*, SIAM J. Matrix Anal. Appl. 36 (2015), 752–774.

[47] 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.