MATHEMATICS OF COMPUTATION
Volume 00, Number 0, Pages 000–000
S 0025-5718(XX)0000-0

STOCHASTIC ALGORITHMS FOR SELF-CONSISTENT CALCULATIONS OF ELECTRONIC STRUCTURES

TAEHEE KO AND XIANTAO LI

Abstract. The convergence property of a stochastic algorithm for the self-consistent field (SCF) calculations of electron structures is studied. The algorithm is formulated by rewriting the electronic charges as a trace/diagonal of a matrix function, which is subsequently expressed as a statistical average. The function is further approximated by using a Krylov subspace approximation. As a result, each SCF iteration only samples one random vector without having to compute all the orbitals. We consider the common practice of SCF iterations with damping and mixing. We prove that the iterates from a general linear mixing scheme converge in a probabilistic sense when the stochastic error has a second finite moment.

1. Introduction

The computation of electronic structures has recently become routine calculations in material science and chemistry [41]. Many software packages have been developed to facilitate these efforts [19, 33, 40, 55]. A central component in modern electronic structure calculations is the self-consistent field (SCF) calculations [41, 48]. The standard procedure is to start with a guessed electron density, and then determine the Hamiltonian, followed by the computation of the eigenvalues and eigenvectors which lead to a new electron density; The procedure continues until the input and output densities are close. Many numerical methods have been proposed to speed up the SCF procedure, see [3, 8, 10, 12, 13, 20, 24, 28, 32, 39, 49, 66–68]. Overall, the SCF still dominates the computation, mainly because of the unfavorable cubic scaling in the computation of the eigenvalue problem. SCF is also a crucial part of ab initio calculations, especially in the Born-Oppenheimer molecular dynamics [43, 61]: The motion of the nuclei causes the external potential to change continuously, and the SCF calculations have to be performed at each time step.

The SCF problem can be formulated as a fixed-point iteration (FPI). One remarkable, but much less explored approach for FPIs, is the random methods [1], which are intimately connected to the stochastic algorithms of Robbins and Monro [14, 53, 64], which in the context of machine learning, has led to the stochastic gradient descent (SGD) methods [7]. The advantage of these stochastic methods is for optimization problems with very large data set, one only calculates a small subset of samples rather than the entire set.

2020 *Mathematics Subject Classification.* Primary MSC 60G52, 65C40.
This work is supported by NSF Grants DMS-1819011 and 1953120.

The main purpose of this paper is to formulate such a stochastic algorithm in the context of SCF, and analyze its convergence property. We first propose to use the diagonal estimator [5] to approximate the matrix function involved in the SCF. The key observation is that with such a diagonal estimator, the approximate fixed-point function can be expressed as a conditional expectation. Consequently, we construct a random algorithm, where we choose a random vector to sample the conditional average. What bridges these two components together is the Krylov subspace method [54] that incorporates the random vector as the starting vector and approximates the matrix function using the Lanczos algorithm. In light of the importance of mixing methods in SCF [3, 8, 24, 28, 32], which often enable and speed up the convergence of the fixed-point iterations, we consider iterative methods with damping and mixing, together with the stochastic algorithm. We also present preliminary numerical results based on a generalized linear mixing scheme.

The convergence of SCF is certainly an outstanding challenge in scientific computing. But our analysis is applicable to general stochastic fixed-point problems. In particular, our convergence analysis treats general stochastic fixed-point problems when the sampling error only has a second moment bound. In this setting, one-step iteration algorithms can be viewed as a Markov chain. Kushner and Yin [34–36] introduced the notion of stochastic stability of discrete-time Markov chains and proved their convergence with probability one when such stability holds. The underlying idea is similar to the Lyapunov function theory for ODEs. The Markov chains considered in [36] have a similar form as the simple mixing scheme. Motivated by their analysis, we shall prove stochastic stability and convergence of the simple mixing scheme. However, such an approach can not be directly extended to general mixing schemes, which correspond to high-order Markov chains. In order to overcome this difficulty, we generalize the Lyapunov functions for extended Markov chains. Remarkably, from this, one can interpret the general mixing scheme as a first-order Markov chain. Further, we establish tools that link the convergence of the extended Markov chains to the properties of the iterates which are of our interest. With the tools and the generalized Lypaunov functions, we will prove that general mixing schemes are also stable and converge to the fixed point with probability one when stability holds, still under the milder condition that the second moment of the stochastic error is finite.

In practice, the models for electronic structure calculations, e.g., the density-functional theory (DFT) [26, 31], have to be discretized in space. One straightforward implementation is the real space discretization using finite difference [4]. To illustrate how to formulate a stochastic fixed-point problem from a more sophisticated discretization, we consider the framework of the self-consistent charge density functional tight-binding method (SCC-DFTB) [19], which has been an important semi-empirical methodology in modern electronic structure simulations. More specifically, Elstner and coworkers devised the method as an improvement of the non-SCC approach. As an application, we will present our stochastic algorithms based on this tight-binding framework, although the application to real-space methods, e.g., [4, 33, 56], is straightforward. In the computational chemistry literature, the stochastic DFT [15] shows resemblance to the present work, especially with the use of randomized algorithms for estimating traces and approximation methods for matrix-vector products. The trace estimator in their work is essentially equivalent to the diagonal estimator [5], which is also used in this work. In [15] the density

matrix is approximated by Chebyshev polynomials, while in our approach, we use the Krylov subspace approximation. More importantly, the framework proposed in [15] is mainly computational. In contrast, this paper presents several theoretical results that are crucial to understanding the performance of such algorithms. Another class of methods that also work with the density matrix is linear-scaling algorithms for DFT [9], which is achieved by exploiting the sparsity of the density matrix.

The rest of the paper is organized as follows. We first present general stochastic fixed-point problems and mixing schemes. Section 3 presents convergence analysis, with emphasis on convergence in probability and the implication to the iteration complexity. In Section 4, we focus specifically on electronic structure calculations. We review the SCF procedure in a tight-binding model [19]. We show that the electronic charges can be expressed as a trace formula. Based on such expressions, we construct a stochastic algorithm from the methods in Section 2 and make a preliminary comparison of the stochastic and deterministic algorithms. In Section 5, we present some numerical results.

2. Stochastic Fixed-Point Iterations and Mixing schemes

We consider numerical methods for the fixed-point problem,

$$q = K(q) = \mathbb{E}[k(q, v)]. \tag{2.1}$$

Here $q \in \mathbb{R}^N$, and the mapping $K : \mathbb{R}^N \mapsto \mathbb{R}^N$, is represented as the expectation of some random mapping $k(q, v)$ with respect to a random vector v whose distribution is known in advance. This problem will be referred to as a *stochastic fixed-point problem*.

A direct approach for the problem (2.1) is the fixed-point iterations,

$$q_{n+1} = K(q_n). \tag{2.2}$$

More generally, one can introduce damping and mixing to improve the convergence, as follows,

$$q_{n+1} = (1 - a_n) \sum_{i=1}^{m} b_i q_{n-m+i} + a_n \sum_{i=1}^{m} b_i K(q_{n-m+i}). \tag{2.3}$$

Here $a_n \in (0, 1)$ can be regarded as a damping parameter; $m \in \mathbb{N}$, and the m constants $b_1, b_2, .., b_m \in \mathbb{R}$ satisfy the conditions that $\sum_{i=1}^{m} b_i = 1$ and $b_i \geq 0$.

Meanwhile, directly implementing the scheme (2.3) for the stochastic fixed-point problem (2.1) requires repeated sampling of $k(q, v)$ [58], which can be computationally demanding. The stochastic algorithm addresses this problem by

$$q_{n+1} = (1 - a_n) \sum_{i=1}^{m} b_i q_{n-m+i} + a_n \sum_{i=1}^{m} b_i k(q_{n-m+i}, v_{n-m+i}). \tag{2.4}$$

Namely, we only sample the random fixed-point function once or a few times in each iteration as shown **Algorithm** 1.

To better describe the linear mixing method (2.4), we denote by

$$B_m(y_n) := \sum_{i=1}^{m} b_i y_{n-m+i}, \tag{2.5}$$

the linear combination of a number of previous iterations. The terminology of *linear* mixing simply means that the right-hand side forms a linear combination of the previous iterates. As outlined in **Algorithm** 1, the implementation is quite straightforward.

Algorithm 1: Linear mixing method

Data: $\{a_n\}$, $\{b_i\}_{i=1}^m$,$\{q_i\}_{i=1}^m$
Result: Approximate fixed-point
for $n = m, m+1, ...,$ *until convergence* **do**
  Compute $k(q_{n+1}, v_{n+1})$;
  $q_{n+1} = (1 - a_n)B_m(q_n) + a_n B_m(k(q_n, v_n))$;
end

When $m = 1$, the linear mixing scheme (2.6) is reduced to

$$q_{n+1} = (1 - a_n)q_n + a_n k(q_n, v_n), \tag{2.6}$$

which is known as simple mixing in electronic structure calculations. Under appropriate assumptions, the iterations from the simple mixing (2.6) have been shown to converge almost surely [1]. Similar results are established for the stochastic gradient descent (SGD), in the machine learning literature [7], which originated with the work [53]. However, an important question is whether the linear mixing method (1) converges with a general depth $m \geq 1$ under mild assumptions. In the following section, we show that despite the random noise within the sample $k(q, v)$, the mixing scheme 1 for any $m \geq 1$ converges in the probabilistic sense.

Remark 2.1. The mixing methods require multiple initial guesses. They can be computed from the simple mixing method (2.6). Alternatively, this can be done by setting $m = 1$ to generate the second iteration q_2, and then $m = 2$ to find q_3, until all the initial vectors are computed, see [59]. For simplicity, we assume that all the m initial vectors have been computed, and our analysis focuses on the subsequent iterations.

Remark 2.2. In practice, there are situations where $K(q)$ is approximated by $K_\ell(q)$, which seems to be easier to compute. The parameter ℓ indicates the order of such an approximation. Compared to the original problem (2.1), this approximation leads to a perturbed fixed-point problem,

$$q = K_\ell(q).$$

In Section 4 we will quantify the effect of such perturbation in the context of electronic structure calculations with Theorem 4.3 and the numerical results in Figure 1.

3. Convergence Theorems and Complexity Estimates

In this section, we present convergence analysis for the linear mixing scheme (1), which is applicable to a large class of stochastic fixed-point problems. Under standard assumptions, these theorems highlight the stochastic stability properties and probabilistic convergence of the mixing algorithm.

Let us first outline the main ingredients in the proofs at the high level. To characterize stochastic stability, we define a family of Lyapunov functions whose input

contains the m iterates and the intermediate stochastic errors. With the results in Appendix D, we will show that the extended Markov chains will produce non-negative supermartingales with the family of Lyapunov functions with some perturbations. In the end, we employ the optional stopping theorem on non-negative supermartingales to complete the proof [36, 51, 63].

3.1. **Assumptions.** The first assumption is that the mapping K is *locally* contractive, since many problems in practice, including electron structure calculations, are not expected to be globally contractive. Remarkably, as we show after Theorem 3.2, this condition can be further relaxed to a stability condition associated with the Jacobian of the mapping K at its fixed-point q^*, under which the following convergence theorems still hold.

Assumption 1. For some $\rho > 0$ and some $c \in (0, 1)$,

$$\|K(x) - K(y)\|_2 \leq c\|x - y\|_2 \tag{3.1}$$

for all $x, y \in \mathcal{B}(q^*, \rho)$ where $\mathcal{B}(q^*, \rho)$ is the ball centered at q^* of radius ρ with respect to the 2-norm.

The second assumption is on the random error arisen in the evaluation $K(q)$.

Assumption 2. For every $q \in \mathcal{B}(q^*, \rho)$, the random error ξ,

$$\xi(q, v) := k(q, v) - K(q), \tag{3.2}$$

satisfies,

$$\mathbb{E}\big[\xi(q, v)|q\big] = 0 \quad \text{and} \quad \sup_{q \in B(q^*, \rho)} \mathbb{E}\big[\|\xi(q, v)\|_2^2|q\big] \leq \Xi. \tag{3.3}$$

The first condition means that $\xi(q, v)$ has zero mean, or the approximation by $k(q, v)$ is unbiased. The second condition states that the variance of the random error is uniformly bounded in the domain of K defined in Assumption 1. These conditions are standard in the machine learning theory [7]. In terms of application, Assumption 2 is fulfilled by the construction of a stochastic algorithm for the DFTB+ as we will explain in Corollary 4.6.

3.2. **Stochastic Stability and Probabilistic Convergence.** We begin by introducing notations for the following theorems. $\mathbb{I}_A$ denotes the indicator function which values one in the event A and zero otherwise. Let R_n be the residual equal to $K(q_n) - q_n$ and $\xi_n := \xi(q_n, v_n)$ be the stochastic error at step n. We define

$$G_n := R_n + \xi_n \tag{3.4}$$

as the sum of the residual and the stochastic error at step n.

Let us first establish the convergence of the simple mixing scheme. Motivated by the analysis in [36], we start by defining a perturbed Lyapunov functional,

$$V_n(q_n) = V(q_n) + \delta V_n, \tag{3.5}$$

where

$$V(q) = \|q - q^*\|_2^2, \quad \delta V_n = \Xi \sum_{i=n}^{\infty} a_i^2. \tag{3.6}$$

Theorem 3.1. *Assume that the damping parameters $\{a_n\}$ satisfy*

$$\sum_n a_n = \infty, \quad \sum_n a_n^2 < \infty, \quad a_n \leq \frac{1-c}{(1+c)^2} \ \ \forall n \in \mathbb{N}. \tag{3.7}$$

Under Assumptions 1 and 2, the simple mixing scheme (2.6) ($m = 1$) *has the following properties:*

(i) *The iterations $\{q_n\}_{n=1}^{\infty}$ leave the ball $B(q^*, \rho)$ with probability,*

$$\mathbb{P}\left\{\sup_n \|e_n\|_2 > \rho | q_1\right\} \mathbb{I}_{\{q_1 \in B(q^*,\rho)\}} \leq \frac{V_1(q_1)}{\rho^2},$$

where the function V_1 is given as (3.5).

Each path $\{q_n\}_{n=1}^{\infty}$ that stays in the ball will be called a stable path.

(ii) *Each stable path converges to q^* with probability one, i.e.,*

$$\mathbb{P}\left\{\lim_{n\to\infty} q_n = q^* | \{q_n\}_{n=1}^{\infty} \subset B(q^*, \rho)\right\} = 1.$$

Consequently, the iterations $\{q_n\}_{n=1}^{\infty}$ converge to q^* with probability at least $1 - \frac{V_1(q_1)}{\rho^2}$.

To highlight the main theoretical results, the proof is included in Appendix E.

To handle the linear mixing scheme (1) with $m \geq 2$, we work with the m vectors at each step. Motivated with the framework and notations in [36], we define an extended state variable by lumping every m iterations coupled with the stochastic noises

$$X_n := (q_{n+m-1}, \xi_{n+m-2}, q_{n+m-2}, .., \xi_n, q_n) \in \mathbb{R}^{(2m-1)N}, \tag{3.8}$$

which forms a first-order Markov chain. In accordance with this, we consider a filtration $\{\mathcal{F}_n\}$ which measures at least $\{X_i, i \leq n\}$. Let us denote by $\mathbb{E}_n$ the expectation conditioned on $\mathcal{F}_n$.

We assume that the mixing coefficient satisfies that

$$\sum_{i=1}^{m} b_i = 1, \quad b_i \geq 0, \quad b_m > 0. \tag{3.9}$$

The general case $m \geq 2$ requires a more sophisticated construction of the Lyapunov function. We define a more general Lyapunov function as

$$\|X_n\|_n = \|e_{n+m-1}\|_2^2 + \sum_{j=2}^{m} \sum_{i=j}^{m} b_{m-i+1} \|e_{n+j-2+m-i} + a_{n+m-3+j} G_{n+j-2+m-i}\|_2^2. \tag{3.10}$$

The subscript n in $\|\cdot\|_n$ is meaningful, since this Lyapunov function depends on n as the coefficient $a_{n+m-3+j}$ in each summand does so.

Together with this function, we define a perturbed Lyapunov function as follows,

$$V_n(X_n) = \|X_n\|_n + \Xi \sum_{i=n}^{\infty} \chi_i, \tag{3.11}$$

where

$$\chi_n := \sum_{j=1}^{m} b_{m-j+1} a_{n+m-2+j}^2. \tag{3.12}$$

Note that χ_n is recognized as a weighted sum of the m squares of the damping parameters. If $m = 1$, it is exactly the case of simple mixing in (3.6).

Theorem 3.2. *Assume that the damping parameters $\{a_n\}$ satisfy* (3.7) *and the mixing coefficient satisfies* (3.9). *Under Assumptions 1 and 2, the general mixing scheme* (2.4) *($m \geq 2$) satisfies*

(i) *The iterations $\{q_n\}_{n=1}^{\infty}$ leave the ball $B(q^*, \rho)$ with probability,*

$$\mathbb{P}\left\{ \sup_{n \geq m+1} \|e_n\|_2 > \rho | \{q_i\}_{i=1}^m \right\} \mathbb{I}_{\{\{q_i\}_{i=1}^m \subset B(q^*,\rho)\}} \leq \frac{V_1(X_1)}{\rho^2},$$

where V_1 is given as (3.11).

(ii) *Each stable path converges to q^* with probability one,*

$$\mathbb{P}\left\{ \lim_{n\to\infty} q_n = q^* | \{q_n\}_{n=1}^{\infty} \subset B(q^*, \rho) \right\} = 1.$$

Consequently, the linear mixing scheme ($m \geq 2$) converges to q^* with probability at least $1 - \frac{V_1(X_1)}{\rho^2}$.

The proof for general m is similar to that of Theorem 3.1. We refer readers to the proof in Appendix E.

The contractive property (3.1) clearly plays an important role in the analysis of fixed-point iterations. But this starting point is not specific to a stochastic algorithm. Rather, it is also essential in deterministic settings [58, 59]. In the context of SCF iterations, [11,39] showed that with a damping term, the contractive property (3.1) of the modified map,

$$K_\theta := (1 - \theta)q + \theta K(q), \tag{3.13}$$

can be guaranteed when the eigenvalues of the Jacobian of the fixed-point map at q^*, here denoted by $K'(q^*)$, are real and less than 1. This implies that there exists a $\theta_{\max} > 0$, such that whenever $0 < \theta < \theta_{\max}$, the spectral radius of $K'_\theta(q^*)$ is less than 1.

We found that it is enough to assume that $K'(q^*)$ has eigenvalues in $\mathbb{C}$ with real part less than 1 to ensure the local contraction (3.1) under some vector norm. This is more general than the conditions in [11, 39], as indicated in the following theorem,

Theorem 3.3. *Suppose that K is continuously differentiable and $\mathrm{Re}(\lambda) < 1$ for each eigenvalue λ of $K'(q^*)$. Then, there exists an unitary transformation U and a vector norm, $\|x\|_D = (x, Dx)^{1/2}$ for some $D \succ 0$, and under this norm the mapping $K_{U,\theta}(p) = (1 - \theta)p + \theta U^{-1}K(Up)$ is locally contractive for any $0 < \theta < \theta_{\max}$.*

The proof can be found in Appendix C.

Remark 3.4. The condition on $K'(q^*)$ in Theorem 3.3 is a general condition that ensures the local contractive property for fixed-point problems. For specific problems, the conditions may be attributed to direct physical interpretations. For instance, in the context of electronic structure calculations, the Jacobian has been factored into matrices $K'(q^*) = AB$, where A and B are symmetric and the eigenvalues of A have the same sign [11,39]. More specifically, in [39], A corresponds to the functional derivative of the sum of the Coulomb potential and the exchange-correlation

potential with respect to the electron density and B can be identified as the independent particle polarizability matrix. In this case, the Jacobian of the fixed-point map is of the form AB [39, Eq. 2.8], and $I - AB$ is the dielectric operator. The authors argued that, in practice, A is approximately positive-definite. Under this assumption, the eigenvalues of AB are all real and $\lambda(AB) < 1$ holds provided that $\lambda(I - AB) > 0$, which can be interpreted as the structural stability [62]. Similarly, the Jacobian of the fixed-point map for the density-matrix in [11] is also expressed as AB (in the proof of Theorem 3.4 in [11]) and $-A^{-1} + B$ is the Hessian of the energy functional at a minimizer [11, according to Eq. 2.6], which is positive-definite. Again, with the assumption that A is negative-definite, justified by the Aufbau principle, it follows that $\lambda(I - AB) > 0$ which implies $\lambda(AB) < 1$.

Now we demonstrate how the results in Theorems 3.1 and 3.2 can still be retained under such a relaxed condition. Without loss of generality, we consider the simple mixing scheme. Choose $\theta > 0$ such that the mapping $K_{U,\theta}$ is contractive according to Theorem 3.3. Let a_n be any damping coefficient such that $\theta > a_n > 0$. Using the unitary transformation U in Theorem 3.3, one can obtain equivalent iterations using the following steps,

$$q_{n+1} = (1 - a_n)q_n + a_n\big(K(q_n) + \xi_n\big), \tag{3.14a}$$

$$p_{n+1} = (1 - a_n)p_n + a_n\big(K_U(p_n) + U^{-1}\xi_n\big), \tag{3.14b}$$

$$p_{n+1} = (1 - \frac{a_n}{\theta})p_n + \frac{a_n}{\theta}\big(K_{U,\theta}(p_n) + \theta U^{-1}\xi_n\big). \tag{3.14c}$$

The first line is simple mixing with a mapping K. By multiplying U^{-1} from the left, the second expression follows by defining $K_U(p) := U^{-1}K(Up)$. More importantly, the third expression can be viewed as the simple mixing scheme with the contraction $K_{U,\theta}$ by Theorem 3.3. Moreover, the constant factor θU^{-1} in front of the error ξ_n does not affect the assumption 2. Therefore, the analysis in Theorems 3.1 and 3.2 can be applied to (3.14c) to obtain the same result in terms of $\{p_n\}$, which can be extended to the iterations $\{q_n\}$ using the equivalence of norms. The same idea can be applied to the general linear mixing scheme (2.4) due to the linear combination of the fixed-point functions.

3.3. **The iteration complexity of the stochastic linear mixing method.** In this section, we study the iteration complexity, that is, the number of required iterations to reach an accuracy threshold ϵ. In terms of stochastic approximation methods, the iteration complexity has been an important topic in optimization problems [7].

For the fixed-point problems (2.1), a direct calculation has been proved to be linearly convergent [11, 39], suggesting that the number of iterations is $\Theta(\log \frac{1}{\epsilon})$. To obtain the corresponding complexity of a stochastic algorithm, we use the same techniques for the previous theorems, and deduce the following inequality for the linear mixing scheme (1),

$$\mathbb{P}\left\{\sup_{m+1\le n\le j} \|e_n\|_2 > \rho | \{q_i\}_{i=1}^m\right\} \mathbb{I}_{\{\{q_i\}_{i=1}^m \subset B(q^*,\rho)\}} \le \frac{V_1(X_1)}{\rho^2}, \tag{3.15}$$

where V_1 is defined as,

$$V_1(X_1) = \|X_1\|_1 + \Xi \sum_{i=1}^{j} \chi_i. \tag{3.16}$$

We denote by $E_j := \{q_n \in B(q^*, \rho),\ \forall n \in \{1, 2, .., j\}\}$, i.e., the event that the first j iterates lie in $B(q^*, \rho)$.

Theorem 3.5. *With the mixing coefficient in* (3.9) *and a non-negative sequence* $\{a_n\}$ *with* $a_n \leq \frac{1-c}{(1+c)^2}$, *the linear mixing scheme* (1) *with initial guesses* $\{q_i\}_{i=1}^m \subset B(q^*, \rho)$ *satisfies that* $\mathbb{P}\{E_j|q_1\} \geq 1 - \frac{V_1(X_1)}{\rho^2}$ *and additionally,*

$$\mathbb{E}\big[\|\bar{q} - q^*\|_2^2 \mathbb{I}_{E_j}|q_1\big] \leq u_j, \tag{3.17}$$

where $\bar{q}$ *is an averaged solution,*

$$\bar{q} := \sum_{n=1}^{j} \left(\frac{A_n}{\sum_{n=1}^{j} A_n}\right) q_n,$$
$$u_j := \frac{V_1(X_1)}{(1-c)\left(\sum_{n=1}^{j} A_n\right)},$$

where A_n *is defined as* $A_n := \sum_{j=1}^{m} b_{m-j+1} a_{n+m-2+j}$ *and* $V_1(X_1)$ *is from* (3.16).

Consequently, the iterations from the linear mixing scheme ($m \geq 2$) *follow the inequality*

$$\mathbb{P}\Big\{\|\bar{q} - q^*\|_2 \leq \epsilon|q_1\Big\} \geq 1 - \frac{V_1(X_1)}{\rho^2} - \frac{u_j}{\epsilon^2}. \tag{3.18}$$

We refer the readers to the proof in E.3.

Now, to obtain a specific complexity estimate, we consider the damping parameter $a_n = \frac{a}{n^\beta}$ for $\beta \in (\frac{1}{2}, 1)$ and the mixing coefficient in (3.9). By applying the integral test to $a_n = \frac{a}{n^\beta}$, one can bound $V_1(X_1)$ as

$$V_1(X_1) \leq \|X_1\|_1 + \Xi \sum_{n=1}^{\infty} \chi_n =: C_{a,\beta} < \infty. \tag{3.19}$$

To proceed, we call q an ϵ-solution if $\|q - q^*\|_2 \leq \epsilon$. By applying the above theorem, we obtain a corollary as follows.

Corollary 3.6. With the mixing coefficient in (3.9), suppose that the damping parameter is given by $a_n = \frac{a}{n^\beta}$ with $\beta \in (\frac{1}{2}, 1)$ and sufficiently small $a > 0$. Then, for any tolerance $\epsilon > 0$ and failure probability $\gamma \in \left(\frac{C_{a,\beta}}{\rho^2}, 1\right)$, with probability at least $1 - \gamma$, the linear mixing scheme finds an ϵ-solution within the number of iterations,

$$j = \Theta\left(\left(\frac{1}{\epsilon^2\left(\gamma - \frac{C_{a,\beta}}{\rho^2}\right)}\right)^{\frac{1}{1-\beta}}\right). \tag{3.20}$$

Θ here is the notation for complexity.

Proof. By letting $\gamma = \frac{V_1(X_1)}{\rho^2} + \frac{u_j}{\epsilon^2}$ in (3.18), using the upper bound $C_{a,\beta}$ for the first term, and observing that $u_j = \mathcal{O}(j^{1-\beta})$ from the integral test, we deduce this result for the averaged solution $\bar{q}$ from Theorem 3.5. □

4. A stochastic framework for a tight-binding approximation of DFT

4.1. **The DFTB+ model.** As a discretization of Density Functional Theory (DFT), tight-binding (TB) approaches have been widely applied for larger electronic systems, particularly due to the fact that they do not require meshes. Among various TB schemes, SCC-DFTB [19] has shown great success for many different molecular and material systems. Part of the success can be attributed to the incorporation of long-range Coulomb interactions. In addition, the implementation allows a self-consistent calculation to determine the charge distribution. Here, we briefly introduce SCC-DFTB [19].

We let M and N, $M \geq N$, be respectively the number of atomic orbitals and nuclei. The atomic orbitals can be naturally labelled by $[M] = \{1, 2, \cdots, M\}$. Let α^j be a multi-index for the atomic orbitals assigned to the j-th atom, i.e., $\alpha^j = \{\alpha_1^j, \alpha_2^j, \cdots, \alpha_{m_j}^j\} \subset [M]$; $\sum_{j=1}^N m_j = M$. For instance, by $\nu \in \alpha^j$ we mean that the atomic orbital ν is associated with the j-th atom. Such notations are particularly useful for a system with multiple species, for which m_j varies. Further, they can be used to indicate those elements in the Hamiltonian matrix H and the overlap matrix S that represent interactions among the atoms, as we explain next. We denote the list of electronic charges associated with the atoms by $q = (q(1), .., q(N))^T \in \mathbb{R}^N$.

The DFTB+ model involves a generalized eigenvalue problem and Hamiltonian corrections using linear response. In particular, the algorithm in [19] finds a solution of charge vector q by iterating through the following equations,

$$\begin{aligned}
&Hc_i = \epsilon_i S c_i \text{ with the eigenpair } (\epsilon_i, c_i),\\
&q(j) = \frac{1}{2}\sum_{i=1}^{M} n_i \sum_{\mu\in\alpha^j}\sum_{\nu=1}^{M}(c_{\mu i}^* c_{\nu i} S_{\mu\nu} + c_{\nu i}^* c_{\mu i} S_{\nu\mu}),\ j = 1, 2, \cdots, N\\
&H_{\mu\nu}^1 = \frac{1}{2} S_{\mu\nu}\sum_{j=1}^{N}(\gamma_{ij} + \gamma_{kj})\Delta q(j),\\
&H = H^0 + H^1,
\end{aligned} \tag{4.1}$$

where

$$n_i = f(\epsilon_i), \quad H_{\mu\nu}^0 = \langle\varphi_\mu|\hat{H}_0|\varphi_\nu\rangle, \quad S_{\mu\nu} = \langle\varphi_\mu|\varphi_\nu\rangle. \tag{4.2}$$

The function f denotes the Fermi-Dirac distribution:

$$f(x) = \frac{2}{1 + \exp(\beta(x - \mu))}, \tag{4.3}$$

with μ being the Fermi energy and β being the inverse temperature. Specifically, $n_i = f(\epsilon_i)$ denotes the occupation number for the energy level ϵ_i.

To explain the notations, here we briefly outline the algorithm in the DFTB model. In the implementation, one starts with a set of preselected localized atomic orbitals $\{\varphi_\mu\}$, the symmetric matrices $H^0 \in \mathbb{R}^{M\times M}$ and $0 \prec S \in \mathbb{R}^{M\times M}$ in (4.2) are defined as the Hamiltonian matrix with the non-SCC TB method [19] and the usual overlap matrix, respectively. In the SCC-DFTB procedure, they are parameterized in terms of the nuclei positions. The first line of (4.1) amounts to a diagonalization of the pair (H, S), with $c_{\nu i}$ denoting the ν-th element of the eigenvector c_i (which corresponds to the atomic orbital ν). The eigenvalues and eigenvectors are then

used to compute the electronic charges q. With the updated charges, one updates the matrix H^1 and the total Hamiltonian before the algorithm enters the next iteration.

In H_1, μ and ν are labels of two atomic orbitals. Since there might be multiple species in the system, they are designated as multi-indices including the orbitals associated with the i-th atom and the k-th atom, respectively. In addition, the coefficient γ_{ij} accounts for the Coulomb interaction between the i-atom and the j-th atom if $i \neq j$ and the self-interaction if $i = j$. In addition, the charge fluctuation $\Delta q(j)$ is defined as $q^0(j) - q(j)$, where $q^0(j)$ is the electronic charge when the atom is in isolation. The formal description on the role of both the quantities γ_{ij} and $\Delta q(j)$ in the DFTB+ model (4.1) is beyond the focus of this paper. For more details, we refer readers to [19].

The steps in (4.1) can be repeated until the charge vector, q, converges. The SCF problem can be reduced to a fixed-point iteration problem (FPI) [39], which for the SCC-DFTB model, can be described as follows. Given q_n as the input, we update the Hamiltonian $H_n = H^0 + H^1(q_n)$ and solve the generalized eigenvalue problem, $H_n U_n = S U_n \Lambda_n$. Using the eigenvalues and eigenvectors, we compute q_{n+1} as the output according to the second equation in (4.1). This procedure can be simplified to a fixed-point iteration,

$$q_{n+1} = K(q_n). \tag{4.4}$$

The mapping K will be expressed as a matrix-vector form (4.10) as we will demonstrate in the next subsection.

After obtaining an approximate limit, the force $F = (F_\alpha) \in \mathbb{R}^N$ can be computed from the total energy E,

$$F_\alpha = -\frac{\partial E}{\partial R_\alpha}, \quad E := \sum_{i=1}^{M} n_i \epsilon_i + E_{rep}, \tag{4.5}$$

where $R = (R_\alpha) \in \mathbb{R}^N$ and E_{rep} describes the repulsion between the nuclei. The calculation of the forces enables geometric optimizations and molecular dynamics simulations [19]. In this paper, we will only focus on the charge iterations. The integration with the force calculation will be addressed in separate works.

A direct implementation of (4.4), however, usually does not lead to a convergent sequence, mainly due to the lack of contractiveness of the mapping K. Practical computations based on (4.4) are often accompanied with a mixing and damping strategy as discussed in [20, 39, 59]. For example, one can use the simple mixing scheme (2.6). More generally, mixing methods [1, 20], which use multiple previous steps, such as the linear mixing (2.4), are commonly employed in practice.

4.2. **Matrix representation for charge functions.** In this section, we present an expression of the charge at an atom in terms of the trace of a matrix. This is an important step towards the construction of stochastic algorithms. A close inspection of the coefficients in the first line of the equation (4.1) reveals the following formula.

Lemma 4.1. *The electronic charge associated with the j-th atom, $q(j)$ in the system* (4.1)*, can be expressed in terms of the trace of a matrix as*

$$q(j) = \operatorname{tr}\left(E_j^T L f(A) L^{-1} E_j\right), \tag{4.6}$$

where $A = L^{-1}HL^{-T}$ with the Cholesky factorization $S = LL^T$. Here $E_j \in \mathbb{R}^{M\times m_j}$ is the rectangular submatrix of the $M \times M$ identity matrix obtained by pulling out the columns according to the multi-index α^j (with dimension denoted by m_j).

Proof. Denote by S_j the rectangular submatrix of the overlap matrix S, with columns associated with indices in α^j. Define $I_j := E_jE_j^T \in \mathbb{R}^{M\times M}$. We use the spectral decomposition

$$(L^T)^{-1}f(A)L^{-1} = \sum_{i=1}^{M} f(\epsilon_i)c_ic_i^T, \tag{4.7}$$

where (c_i, ϵ_i) is the eigenpair defined in (4.1).

First, we can rewrite the first equation in (4.1) as follows

$$q(j) = \frac{1}{2}\sum_{i=1}^{M} n_i \sum_{\mu\in\alpha^j}\sum_{\nu=1}^{M}(c_{\mu i}^T c_{\nu i} S_{\mu\nu} + c_{\nu i}^T c_{\mu i} S_{\nu\mu}) = \frac{1}{2}\sum_{i=1}^{M} n_i(c_i^T I_j S c_i + c_i^T S I_j c_i).$$

By using the commutative property of the trace, the first term in the summand can be rewritten as follows

$$c_i^T I_j S c_i = \mathrm{tr}(c_i^T I_j S c_i) = \mathrm{tr}(c_i c_i^T I_j S) = \mathrm{tr}(c_i c_i^T E_j E_j^T S) = \mathrm{tr}(c_i c_i^T E_j S_j^T).$$

By a similar treatment of the second summand, we can rewrite the charge $q(j)$

$$\begin{aligned} q(j) &= \frac{1}{2}\sum_{i=1}^{M} n_i \mathrm{tr}(c_i c_i^T (E_j S_j^T + S_j E_j^T)) \\ &= \frac{1}{2}\mathrm{tr}(L^{-T}f(A)L^{-1}(E_j S_j^T + S_j E_j^T)) \\ &= \mathrm{tr}(E_j^T (L^T)^{-1} f(A) L^{-1} S_j) = \mathrm{tr}(E_j^T L^{-T} f(A) L^T E_j) = \mathrm{tr}(E_j^T L f(A) L^{-1} E_j). \end{aligned}$$

The second equality holds by the spectral decomposition shown above. In the last line, the identity $L^{-1}S_j = L^T E_j$ is used. □

We now turn to the third equation in (4.1), which updates the Hamiltonian matrix at each iteration in the SCC-DFTB procedure. The equation is given in terms of the Hamiltonian H and the overlap matrix S. However, as shown in the preceding lemma, the matrix of the specific form $L^{-1}HL^{-T}$ is required to update the charge q_α. This leads to a reformulation of the third equation in (4.1). Here, we introduce notations as follows: The symbol sym stands for the symmetrization $\mathrm{sym}(A) = A + A^T$. In addition, we define $e \otimes_N v$ with $v = (v_1, v_2, .., v_N)^T$ as follows,

$$e \otimes_N v := (\underbrace{v_1, \ldots, v_1}_{m_1}, \ldots, \underbrace{v_N, \ldots, v_N}_{m_N})^T,$$

where m_j, the number of atomic orbitals associated with the j-th atom, indicates the number of times the element v_j is repeated. As opposed to the Kronecker product notation $\otimes$, this operation copies each element of the vector v as many times as the corresponding index α^j.

Lemma 4.2. *The third equation in* (4.1) *has the following alternative expression,*

$$A = A_0 + \frac{1}{2}\mathrm{sym}\big(L^{-1}\mathrm{diag}(e \otimes_N \Gamma\Delta q)L\big), \tag{4.8}$$

where $A_0 = L^{-1}H_0L^{-T}$ and $\Gamma := (\gamma_{ij}) \in \mathbb{R}^{N\times N}$.

We prove this lemma as follows.

Proof. In the third line of (4.1), the correction term can be rewritten as

$$\begin{aligned} H^1_{\mu\nu} &= \frac{1}{2}S_{\mu\nu}\sum_{j=1}^{N}(\gamma_{ij}+\gamma_{kj})\Delta q(j), \quad \Delta q := (\Delta q(1),...,\Delta q(N))^T \\ &= \frac{1}{2}S_{\mu\nu}\Big(\Gamma_i^T\Delta q + \Gamma_k^T\Delta q\Big), \quad \Gamma_i \text{ is the } i\text{-th row vector of } \Gamma \\ &= \frac{1}{2}S_{\mu\nu}\Big(i-\text{th entry of } \Gamma\Delta q + k-\text{th entry of } \Gamma\Delta q\Big) \end{aligned}$$

$$\Longrightarrow H^1 = \frac{1}{2}(\underbrace{\mathrm{diag}(e\otimes_N \Gamma\Delta q)S}_{\text{row operation by } \mu} + \underbrace{S\mathrm{diag}(e\otimes_N \Gamma\Delta q)}_{\text{column operation by } \nu}).$$

The desired expression is obtained by multiplying L^{-1} to left and L^{-T} to right. □

In summary, the system (4.1) can be concisely rewritten as

$$\begin{aligned} q(j) &= \mathrm{tr}(E_j^T L f(A) L^{-1} E_j), \\ A &= A_0 + \frac{1}{2}\mathrm{sym}(L^{-1}\mathrm{diag}(e\otimes_N \Gamma\Delta q)L). \end{aligned} \tag{4.9}$$

We note that the generalized eigenvalue problem in the system (4.1) is incorporated in the system (4.9) implicitly. Especially based on the charge $q(j)$ in the system 4.9, the mapping K in (4.4) can be explicitly formulated as follows,

$$K:\mathbb{R}^N\to\mathbb{R}^N, \quad K(q) = \begin{bmatrix} \mathrm{tr}(E_1^T L f(A) L^{-1} E_1) \\ \mathrm{tr}(E_2^T L f(A) L^{-1} E_2) \\ \vdots \\ \mathrm{tr}(E_N^T L f(A) L^{-1} E_N) \end{bmatrix}, \tag{4.10}$$

where N denotes the number of nuclei. We recall that the matrix A in the right-hand side of (4.9) involves the charge vector q within the term Δq, which means that K is a mapping of q.

4.3. **A stochastic algorithm for the DFTB+ model.** According to (4.10), one can directly calculate $K(q)$ when the diagonal of the matrix $Lf(A)L^{-1}$ is explicitly known, while it requires the eigen decomposition of A, which is expensive for large matrices. Alternatively, we employ the diagonal estimator A.1 (see similar applications to electronic structure calculations [5, 57]). Within this diagonal estimator, one has to compute a matrix-vector product, in our case, $f(A)L^{-1}v$, which requires the eigen decomposition of A again. To bypass a full diagonalization, we use the Krylov subspace method [17, 54] to approximate the matrix-vector product. This method yields a fairly good approximation for sparse matrices. Error estimates of the Krylov approximation have been proposed in [17, 18, 54] for the case of the exponential-like functions. However, the Fermi-Dirac distribution (4.3) clearly does not belong to this family of functions, and an error estimate requires a different proof.

Theorem 4.3. *Suppose that A is a symmetric matrix and $f(x)$ is the Fermi-Dirac distribution in* (4.3)*. Then, for any integer $\ell > s$, the error of the Krylov subspace method can be bounded by,*

$$\left\| f(A)v - \|v\|_2 V_\ell f(T_\ell) e_1 \right\|_2 \leq \frac{4M(\rho)\|v\|_2 \rho^{-\ell}}{\rho - 1}, \tag{4.11}$$

where V is the total variation of $f^{(s)}(x)$. The constants $M(\rho)$ and $\rho > 1$ depend on $f(x)$ and the difference between the smallest and the largest eigenvalues of A. Consequently, as the degree ℓ increases, one expects the accuracy from the Krylov approximation to improve, namely,

$$\lim_{\ell \to \infty} \|v\|_2 V_\ell f(T_\ell) e_1 = f(A)v.$$

The proof of this theorem, using tools from spectral approximations in the previous works [60, 65], is given in Appendix B.

Remark 4.4. The theorem still holds for any continuously differentiable function that can be extended analytically to some Bernstein ellipse according to results in [60]. Furthermore, this shows that the Krylov subspace approximation with such a function improves the error bound for the Chebyshev approximation by noticing the inequality (B.1). To be specific, the error bound with Chebyshev approximation decays in a polynomial order of ℓ, but the Krylov subspace approximation has an exponential decaying error bound.

Now, by combining the diagonal estimator (A.1) and the Krylov subspace approximation (4.3), we estimate the diagonal of the matrix $Lf(A)L^{-1}$ as follows,

$$\operatorname{diag}(Lf(A)L^{-1}) = \operatorname{diag}(\mathbb{E}[Lf(A)L^{-1}vv^T]) \approx \operatorname{diag}(\mathbb{E}[\|L^{-1}v\|_2 L V_\ell f(T_\ell) e_1 v^T]), \tag{4.12}$$

where $V_\ell \in \mathbb{R}^{M \times \ell}$ is the left-orthogonal matrix, $T_\ell \in \mathbb{R}^{\ell \times \ell}$ is the tridiagonal matrix from the Lanczos method of ℓ steps and v is a random vector whose covariance is the identity matrix. Here, e_1 is the first standard basis vector in $\mathbb{R}^\ell$.

Especially, for each ℓ, the approximation (4.12) is reduced to the relation

$$K(q) \approx K_\ell(q) = \mathbb{E}[k_\ell(q, v)], \tag{4.13}$$

where the expectation is taken over the random vector v and the random mapping $k_\ell(q, v)$ is defined similar to (4.10),

$$k_\ell(q, v) = \|L^{-1}v\|_2 \begin{bmatrix} \operatorname{tr}(E_1^T L V_\ell f(T_\ell) e_1 v^T E_1) \\ \operatorname{tr}(E_2^T L V_\ell f(T_\ell) e_1 v^T E_2) \\ \vdots \\ \operatorname{tr}(E_N^T L V_\ell f(T_\ell) e_1 v^T E_N) \end{bmatrix}. \tag{4.14}$$

We note that the average $K_\ell(q)$ is not equal to the original mapping $K(q)$ because of the error from the subspace approximation method (4.11). In other words, this yields an approximate fixed-point problem. Nevertheless, we can make the approximation error negligible by selecting a sufficiently large ℓ, as shown in Theorem 4.3.

After obtaining the output $[\|v\|_2, V_\ell, T_\ell]$ from the Lanczos method, an eigensolver should be implemented for the eigen decompostion of T_ℓ in order to perform the Krylov subspace approximation with $f(T_\ell) = U_\ell f(D_\ell) U_\ell^T$. As compared to the original system, this is a much smaller matrix and the computation is much easier.

Algorithm 2: Stochastic Lanczos method for the charge function in (4.10)

Data: A, L, ℓ, n_{vec}
Result: $\frac{1}{n_{vec}}\sum_{i=1}^{n_{vec}} k_\ell(q, v_i)$, Approximation of $K(q)$
Samples $v_1, v_2, ..., v_{n_{vec}}$
Define $V_1 = [v_1, v_2, ..., v_{n_{vec}}]$
$V_2 = L^{-1}V_1$
for $i = 1 : n_{vec}$ **do**
  $v = V_2(:, i)$;
  $[\|v\|_2, V_\ell, T_\ell] = \text{Lanczos}(A, v, \ell)$;
  $V_2(:, i) = \|v\|_2 V_\ell f(T_\ell)e_1$, Krylov subspace approximation for $f(A)v$;
end
$V_2 = LV_2$, Approximation of the matrix $Lf(A)L^{-1}V_1$
Compute the average $\frac{1}{n_{vec}}\sum_{i=1}^{n_{vec}} \text{diag}(V_2(:, i)V_1(:, i)^T) \approx \text{diag}(Lf(A)L^{-1})$
Compute the Monte-Carlo sum, $\frac{1}{n_{vec}}\sum_{i=1}^{n_{vec}} k_\ell(q, v_i)$ using (4.14)

Finally, by incorporating Algorithms 1 and 2 into the system (4.9), we arrive at a stochastic self-consistent algorithm for the DFTB+ model outlined in **Algorithm** 3.

Algorithm 3: Stochastic Self-Consistent Calculation for the DFTB

Data: $\{a_n\}$, $\{b_i\}_{i=1}^m$,$\{q_i\}_{i=1}^m$, $A_0 = L^{-1}H_0L^{-T}$, ℓ, n_{vec}
Result: Approximate fixed-point
for $n = m, m + 1, ...,$ *until convergence* **do**
  $q_{n+1} = (1 - a_n)B_m(q_n) + a_nB_m(k(q_n, v_n))$;
  $A_{n+1} = A_0 + \frac{1}{2}\text{sym}(L^{-1}\text{diag}(e \otimes_N \Gamma\Delta q_{n+1})L)$;
  $k(q_{n+1}, v_{n+1}) = \text{StoLan}(A_{n+1}, L, \ell, n_{vec})$, implement **Algorithm** 2 ;
end

Remark 4.5. In the computation of the stochastic function $k(q_n, v_n)$, we have assumed that the Fermi level μ is given. In the stochastic algorithm framework, this can be done very efficiently using the trace estimator [38, 65] for the density of states, which can be subsequently used to estimate the Fermi level.

We denote the stochastic noise from the relation (4.13) by

$$\xi_\ell(q, v) = k_\ell(q, v) - K_\ell(q). \tag{4.15}$$

By applying Theorem 4.3, we obtain the following result,

Corollary 4.6. The stochastic noise (4.15) has zero mean and a bounded variance in the ball $\mathcal{B}(q^*, \rho)$.

Proof. This is because $K(q)$ is continuous and Theorem 4.3 guarantees the boundedness of the approximation (4.12). In addition, by the relation (4.13), the stochastic noise has zero mean. □

Remark 4.7. In sharp contrast to the deterministic counterpart (2.6), the term $k(q_n, v_n)$ in (3) emphasizes the point that the quantity is only sampled once, motivated by the remarkable success of the stochastic algorithms [53]. This leads

to a significant reduction of one iteration cost and a potential application of the stochastic framework for large-scale systems.

4.4. **A Preliminary Comparison of Stochastic and Direct SCC-DFTB.** In the implementation of stochastic approximation methods, three sources of error arise: *approximation, estimation* and *optimization* [6]. In our case, due to the diagonal estimator (A.1) in which the true distribution of the random vector v is determined by users, we do not consider an estimation error. Rather, we focus on the approximation error and the optimization error.

To be precise, we aim at estimating how close some iterate q obtained from a stochastic algorithm is to a solution q^*,

$$\|q - q^*\|_2 \leq \underbrace{\|q - q_\ell^*\|_2}_{\text{optimization error}} + \underbrace{\|q_\ell^* - q^*\|_2}_{\text{approximation error}}, \tag{4.16}$$

where q_ℓ^* and q^* are the fixed points of the mapping K_ℓ in (4.13) and the exact mapping K in (4.10), respectively.

To quantify the approximation error, we use Theorem 4.3. We assume that K (4.10) satisfies the stability condition in Theorem 3.3 and the same for K_ℓ (4.13) for sufficiently large ℓ based on Theorem 4.3. As we discussed in Theorem 3.3, these mappings can be transformed into contractions with a small auxiliary parameter $\theta > 0$. Therefore, without loss of generality, we assume that K and K_ℓ satisfy the contractiveness in **Assumption** 1 throughout the following analysis.

First, we derive an error bound for the approximation. By definitions of the solutions q^* and q_ℓ^*, we have

$$q_\ell^* - q^* = K_\ell(q_\ell^*) - K_\ell(q^*) + K_\ell(q^*) - K(q^*).$$

By using the triangle inequality, one has,

$$\|q_\ell^* - q^*\|_2 \leq \|K_\ell(q_\ell^*) - K_\ell(q^*)\|_2 + \|K_\ell(q^*) - K(q^*)\|_2 \leq c_\ell \|q_\ell^* - q^*\|_2 + C(\rho)\rho^{-\ell},$$

where $c_\ell \in (0, 1)$ is associated with the contraction K_ℓ. To derive the last term $C(\rho)\rho^{-\ell}$, we recall the relation (4.12) with $K(q)$ in (4.10) and $K_\ell(q)$ in (4.13). In the case of the Hutchinson estimator (A.1) where the sample space consists of a finite number of random vectors of the same length, we apply Theorem 4.3 to (4.12),

$$\begin{aligned}
&\|\mathbb{E}[Lf(A)L^{-1}vv^T] - \mathbb{E}[\|L^{-1}v\|_2 L V_\ell f(T_\ell)e_1 v^T]\|_2 \\
&\leq \|\mathbb{E}[L\left(f(A)L^{-1}v - \|L^{-1}v\|_2 V_\ell f(T_\ell)e_1\right)v^T]\|_2 \\
&\leq \mathbb{E}[\|L\|_2 \cdot \|f(A)L^{-1}v - \|L^{-1}v\|_2 V_\ell f(T_\ell)e_1\|_2 \cdot \|v^T\|_2] \\
&\leq \|L\|_2 \|v\|_2 \frac{4M(\rho)\|L^{-1}v\|_2 \rho^{-\ell}}{\rho - 1}.
\end{aligned}$$

This can be used to estimate the difference $\|K_\ell(q^*) - K(q^*)\|_2$ by noticing the definitions $K(q)$ in (4.10) and $K_\ell(q)$ in (4.13). Overall, the approximation error is estimated as

$$\|q_\ell^* - q^*\|_2 \leq \frac{C(\rho)\rho^{-\ell}}{1 - c_\ell}. \tag{4.17}$$

Now we turn to the optimization error based on the result in Section 3.3. In particular, this can be regarded as a route to compare the stochastic and direct methods. In computing the Mulliken charge $q(j)$ in the system 4.1, the computation involved in the eigenvalue problem scales $\mathcal{O}(M^3)$. In contrast, this only scales

Algorithm	Cost of one iteration	Iterations to reach ϵ	Total cost
Stochastic	$\mathcal{O}(\ell M^2)$	$\mathcal{O}\left(\epsilon^{-\frac{2}{1-\beta}}\right)$	$\mathcal{O}\left(\ell M^2 \epsilon^{-\frac{2}{1-\beta}}\right)$
Exact	$\mathcal{O}(M^3)$	$\Theta\left(\log \frac{1}{\epsilon}\right)$	$\mathcal{O}\left(M^3 \log \frac{1}{\epsilon}\right)$

TABLE 1. Comparison between stochastic and direct algorithms (4.4) in terms of optimization error with tolerance ϵ. Here M is the dimension of the Hamiltonian or overlap matrix.

$\mathcal{O}(\ell M^2)$ within the stochastic Lanczos method 2. The cost for the rest of the procedure in both methods is negligible due to the sparsity of the matrices. This observation leads to the comparison in Table (1).

We are now in a better position to compare the direct method (4.4) and the stochastic method (3). To reach accuracy ϵ in (4.16), Table (1) implies that the direct method requires the following time

$$\mathcal{O}\left(M^3 \log \frac{1}{\epsilon}\right),$$

whereas the stochastic method requires the time, for $\beta \in (\frac{1}{2}, 1)$ in 3.6,

$$\mathcal{O}\left(\ell M^2 \widetilde{\epsilon}^{-\frac{2}{1-\beta}}\right),$$

where

$$\widetilde{\epsilon} := \epsilon - \frac{C(\rho)\rho^{-\ell}}{1 - c_\ell}$$

by using the estimate (4.17).

Therefore, the stochastic method could become advantageous when the number of orbitals, M, is larger than

$$M = \mathcal{O}\left(\ell \left(\widetilde{\epsilon}^{\frac{2}{1-\beta}} \log \frac{1}{\epsilon}\right)^{-1}\right), \tag{4.18}$$

This was motivated by the previous work [6] to compare stochastic optimization methods with direct counterparts.

Remark 4.8. A closer inspection of the Krylov subspace approximation (4.3) implies that the temperature influences the quality of this approximation. Especially, at a low temperature $T \ll 1$, as we discuss after Theorem B.2, the minor axis of a Bernstein ellipse should be small enough to guarantee a reasonable approximation. More specifically, by the definition of the Bernstein ellipse [60], an optimal radius ρ can be found from the following equation

$$\rho - \frac{1}{\rho} = \frac{4\pi k_B T}{\lambda_{\max}(A) - \lambda_{\min}(A)} =: c(T), \tag{4.19}$$

which yields that when $T \ll 1$

$$\rho = \frac{c(T) + \sqrt{c(T)^2 + 4}}{2} \approx 1 + \frac{4c(T) + c(T)^2}{8} \approx 1 + \frac{c(T)}{2}.$$

Therefore, at the low temperature, one can deduce that the approximation error of (4.16) decays exponentially with ℓ roughly as

$$\left(1 + \frac{2\pi k_B T}{\lambda_{\max}(A) - \lambda_{\min}(A)}\right)^{-\ell}.$$

Due to the factor T in this expression, a sufficiently large ℓ should be chosen to obtain a reasonable approximation. In contrast, in the regime of high temperature, the Krylov approximation method becomes very effective, since a large value of ρ can be selected from (4.19).

In summary, the stochastic algorithm (3) offers a new framework for electronic structure calculations. One immediate question is when it is more efficient than a direct SCF method for specific applications, e.g., biomolecules or crystalline solids. It is still a complex issue and a much more comprehensive study is needed to take into account many factors, e.g., parallelization, implementations with sparse matrix factorizations, choosing optimal mixing parameters, variance reduction techniques, etc. We leave these issues to future studies.

5. Numerical Results

In this section, we present preliminary results from some numerical experiments. We consider a system of graphene with 800 atoms. We have chosen the lattice spacing to be 1.4203 Å. In the function f (4.3), we set the Fermi level to be -0.1648 and $\beta = 1052.58$, which corresponds to 300 Kelvin. The Hamiltonian and overlap matrices, together with the matrix Γ are all obtained from DFTB+ [19]. The dimension of these matrices is 3200×3200.

As a reference, the solution q^* of (4.4) is first computed using the simple mixing scheme with damping $a = 0.001$. In addition, upon convergence, we used a centered difference method with step size $h = 0.001$, and computed the Jacobian $K'(q^*)$. We found that all the eigenvalues are real and lie between -6.8859 and -0.1348. In light of Theorem 3.3, the contraction assumption (3.1) is fulfilled under some norm by choosing a small step size a.

Since the original fixed point problem $q = K(q)$ has been replaced by $q = K_\ell(q)$, we first examine the error between the fixed points from those mappings. Figure 1 shows how this error depends on the dimension ℓ of the subspace. The error in the figure is measured in $\|\cdot\|_\infty$ norm and the norm of q^* is around 4. One can observe that the error decreases when the subspace is expanded.

For the rest of the discussions, we choose $\ell = 20$, and we regard the fixed point of $K_{20}(q)$ as the true solution q^*.

We test linear mixing methods (**Algorithm** 3). We pick uniform mixing parameters, i.e., $b_i = 1/m$. In addition, we choose the damping parameter, $a_n = \min\{(50 + 2n)^{-1}, 0.005\}$, which fulfills the conditions in the convergence theorem. The error from 30,000 iterations are shown in Figure 2. To mimic the mean error, we averaged the error over every 1,000 iterations. In addition, we run all the cases with simple mixing for 2,000 iterations, followed with the mixing schemes turned on, to allow these cases to follow the same initial period. Surprisingly, the linear mixing scheme does not seem to have faster convergence than the simple mixing. To further test the convergence, we choose the damping parameter as follows, $a_n = \min\{[50 + 4n^{3/4}]^{-1}, 0.005\}$, and show the results in Figure 3. Interestingly, with this choice of the damping parameter, using more mixing steps (larger m) yields faster convergence.

As a simple exposition, we applied the Anderson mixing method [59] to the stochastic algorithm (3). Note that the mixing coefficient b_n is determined on the fly via a least squares procedure [59]. Figure 4 displays the error from 30,000 iterations of the Anderson's method with $m = 2, 3, 4$ and 5. In the implementations, we choose

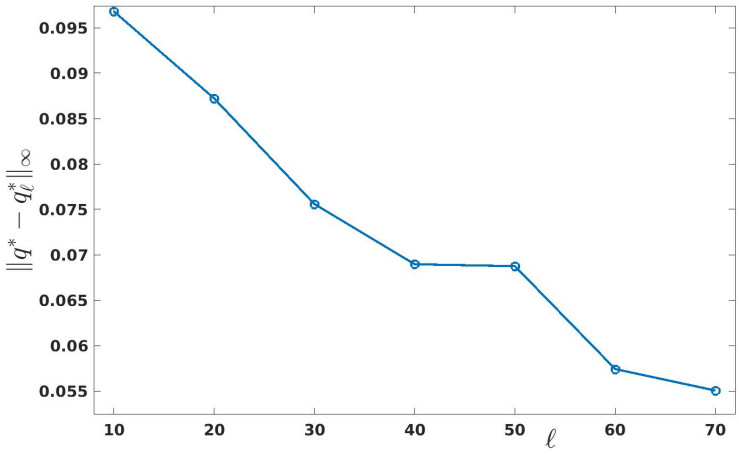

FIGURE 1. The error of the subspace approximation K_ℓ: q^* and q_ℓ^* are respectively the fixed-points of K and K_ℓ. The error is measured in the infinity norm.

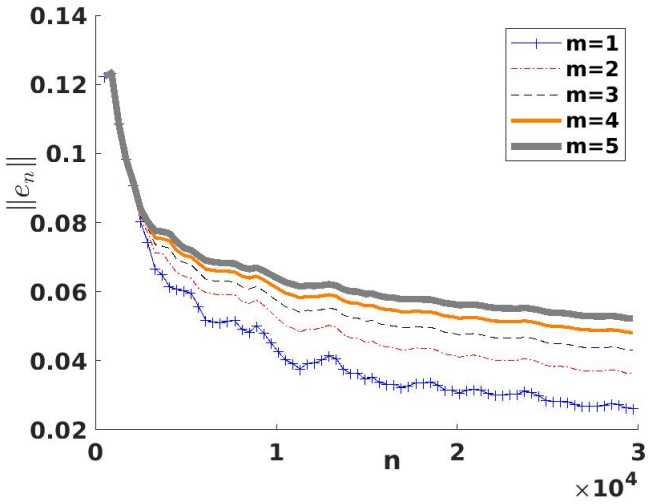

FIGURE 2. The error $\|\bar{q}_n - q^*\|_\infty$ from the linear mixing method (Algorithm 3) with $m = 2, 3, 4, 5$ and 6 using $a_n = \min\{(50+2n)^{-1}, 0.005\}$.

$a_n = [50 + 4n^{3/4}]^{-1}$. Again, due to the stochastic nature, we define the error to be $\bar{q}_n - q^*$ with $\bar{q}_n$ being a local average over the previous 1,000 iterations. The error is then measured by the ∞-norm. One finds that the Anderson mixing does improve the convergence. But the improvement does not seem to be overwhelming.

Figure 5 plots the mixing coefficients b_i's obtained from the Anderson mixing method applied to the algorithm 2 with $m = 3$. It can be observed that these

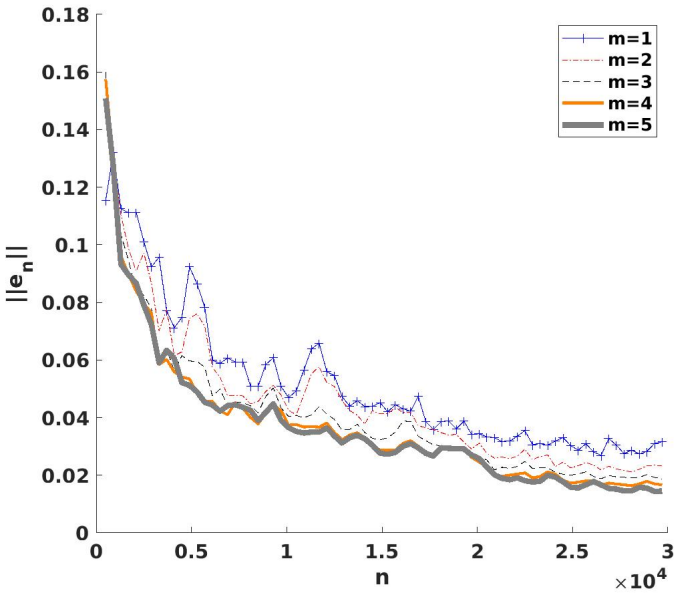

FIGURE 3. The error $\|\bar{q}_n - q^*\|_\infty$ from the linear mixing method (Algorithm 3) with $m = 2, 3, 4, 5$ and 6 using $a_n = \min\{[50 + 4n^{3/4}]^{-1}, 0.005\}$.

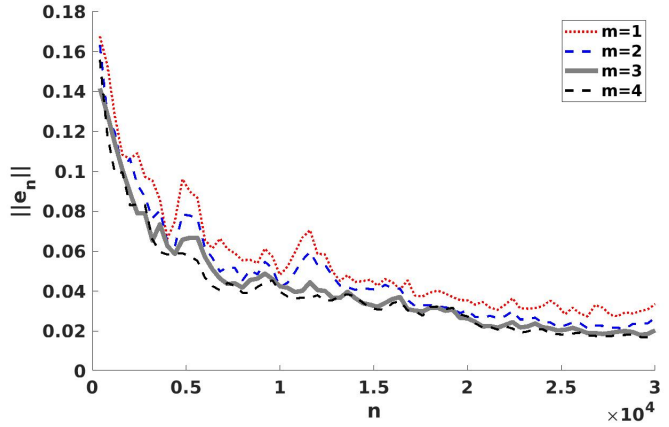

FIGURE 4. The error $\|\bar{q}_n - q^*\|_\infty$ from the Anderson mixing method with $m = 2, 3, 4$ and 5 using $a_n = [50 + 4n^{3/4}]^{-1}$.

coefficients are stochastic in nature. Remarkably, after a short burn-in period, these coefficients tend to fluctuate around the uniform mean $1/m$. One interpretation is that as the iterates q_n get closer to the q^*, the residual error $G(q)$ in (3.4) is dominated by the stochastic error ξ_n. In this case the least-square problem is

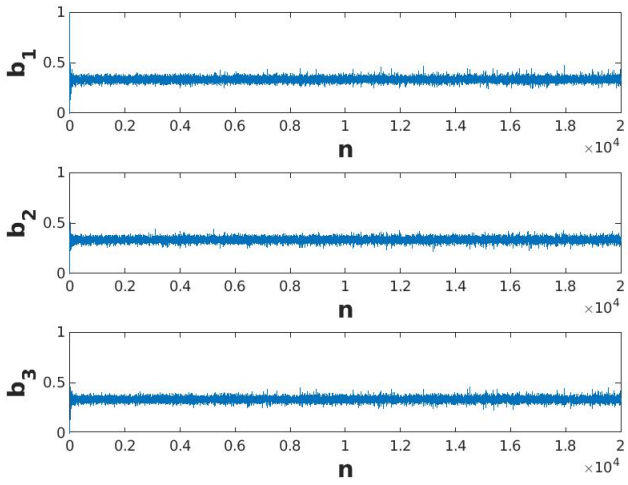

FIGURE 5. The coefficient b_i from the Anderson ($m = 3$) method

mostly determined by the stochastic noise, and it does not show bias toward a particular step.

Lastly, all the computations were performed in Matlab R2020b, with parameters extracted from DFTB [19]. To follow up the discussion in Section 3.3, with 800 atoms, each stochastic iteration takes CPU time 4.75 (seconds), while a direct method takes 52.88 s. When the system size is increased to 1600 atoms, the respective CPU time is 10.12 s and 493.42 s.

6. Conclusion

This paper is motivated by the observation that the main roadblock for extending electronic structure calculations to large systems is the SCF and the full diagonalizations that are involved in each step of the procedure. This observation, for instance, has motivated a great deal of effort to develop linear or sublinear-scaling algorithms that do not directly rely on direct eigevalue computations [9, 21, 22]. Meanwhile, stochastic algorithms have shown promising capability to handle linear and nonlinear problems in numerical linear algebra [42], and computational chemistry [23, 25, 27, 45, 52]. This paper takes an initial step toward a stochastic implementation of the SCF. The main purpose is to establish certain convergence results. In particular, we showed that when the damping parameters are selected a priori, the damped method converges with probability one when the iteration is stable. Additionally, we derived the concentration inequality for the stochastic method. Some of these results are similar to those from the stochastic gradient descent methods in machine learning [7, 16, 29, 46]. A crucial issue in the current approach is the stability: Since the contractive property only holds in the vicinity of the solution, one must establish the stability of the iterations before proving the convergence.

While the convergence is a critical issue, many practical aspects remain as open issues, and they were not studied in this paper. First, how to choose the mixing

parameters b_i's in advance still remains open. Although we have shown the dependence of the error bound on m and the mixing parameters, our analysis does not provide a clear criterion. Secondly, the choice of the damping parameter has a direct impact on the convergence. It would be of practical importance to be able to adjust them on-the-fly, as studied in the machine learning literature [7]. Finally, as discussed in Section 3.3, comprehensive studies are needed to compare the stochastic algorithms to direct implementations of SCF to evaluate the performance for different physical systems. On the other hand, the two approaches do not have to be mutually exclusive in practice. For instance, one can run iterations using stochastic algorithms first, and later switch to a direct method to improve the accuracy at the final stage. Such a strategy is used in machine learning, i.e., the stochastic variance reduction gradient [29]. For example, Reddi et al. [50] showed that alternating between stochastic and direct methods can be better than using only direct methods.

7. Acknowledgment

The authors thank Prof. Kieron Burke and Prof. Lin Lin for discussions and references related to this work. We also thank an anonymous referee for comments regarding Theorem 3.3.

Appendix A. The diagonal estimator

We restate the stochastic framework [5] and the property of the Hutchinson estimator [2] in the following lemma.

Lemma A.1. *For each matrix $A \in \mathbb{R}^{M\times M}$, the follow identity holds,*

$$\operatorname{diag}(A) = \operatorname{diag}(\mathbb{E}[Avv^T]), \tag{A.1}$$

where $v \in \mathbb{R}^M$ is a random vector satisfying,

$$\mathbb{E}[vv^T] = I_{M\times M}. \tag{A.2}$$

Moreover, if the Hutchinson estimator is used, then

$$\operatorname{Var}(\operatorname{diag}[Avv^T]) = \|A\|_F^2 - \sum A_{ii}^2,$$

where the entries of v are i.i.d Rademacher random variables,

$$\mathbb{P}\{v^{(i)} = \pm 1\} = \frac{1}{2}.$$

The *diagonal* can be estimated using a Monte-Carlo sum with n_{vec} i.i.d. random vectors. A variety of such estimators are investigated in [2].

Appendix B. An error estimate on the Krylov subspace approximation

Let $\ell' > \ell > 1$ and $p_{\ell'}(x)$ be the Chebyshev polynomial approximation of degree ℓ' to $f(x)$. We recall that ℓ is the number of iterations for using the Lanczos algorithm. By the triangle inequality, we split the error into three terms,

$$\begin{aligned}\|f(A)v - \|v\|_2 V_\ell f(T_\ell)e_1\|_2 &\le \|f(A)v - p_{\ell'}(A)v\|_2 + \|p_{\ell'}(A)v - \|v\|_2 V_\ell p_{\ell'}(T_\ell)e_1\|_2 \\ &+ \|\|v\|_2 V_\ell p_{\ell'}(T_\ell)e_1 - \|v\|_2 V_\ell f(T_\ell)e_1\|_2.\end{aligned}$$

We will derive upper bounds for these three terms. For the first and third terms, we use Theorem 7.2 in [60]. Meanwhile, we will Theorem 8.1 in [60] to find an upper bound for the second term.

Theorem B.1 (Theorem 7.2 in [60]). *For an integer $s \geq 1$, let f and its derivatives through $f^{(s-1)}$ be absolutely continuous on $[-1,1]$ and suppose that the sth order derivative $f^{(s)}$ is of bounded variation V. Then, for any $\ell > s$, the Chebyshev approximation of degree ℓ, p_ℓ, satisfies,*

$$\|f - p_\ell\| \leq \frac{2V}{\pi s(\ell - s)^s},$$

where $\|h\|$ denotes the supremum norm of the function h.

For a general symmetric matrix A whose spectrum is not necessarily contained in $[-1,1]$, a linear transformation is first applied to A to shift the spectrum to the interval $[-1,1]$. This can be achieved using the linear transformation

$$A \mapsto \tilde{A} := \frac{2A}{b-a} - \frac{b+a}{b-a} I, \quad b := \lambda_{\max},\ a := \lambda_{\min}.$$

With this transformation, we have,

$$f(x) \mapsto \tilde{f}(x) := f(\frac{b-a}{2}x + \frac{a+b}{2}) \approx \tilde{p}(x) \mapsto p(x) := \tilde{p}(\frac{2x}{b-a} - \frac{a+b}{b-a}), \tag{B.1}$$

which means that $p(x)$ is the Chebyshev approximation of $f(x)$ defined on the desired interval. Following this the procedure, the variation of $\tilde{f}^{(s)}(x)$ is proportional to that of $f^{(s)}(x)$ as follows

$$\tilde{V} = \left(\frac{b-a}{2}\right)^s V,$$

where V is the variation of $f^{(s)}(x)$.

Since A is symmetric as defined in (4.8), a direct application of the above theorem yields,

$$\|f(A) - p_{\ell'}(A)\| = \max_{\lambda \in \sigma(A)} |f(\lambda) - p_{\ell'}(\lambda)| \leq \|f - p_{\ell'}\| = \|\tilde{f} - \tilde{p}_{\ell'}\| \leq \frac{V}{2^{s-1}\pi s}\left(\frac{b-a}{\ell' - s}\right)^s.$$

Consequently, we have

$$\|f(A)v - p_{\ell'}(A)v\|_2 \leq \frac{\|v\|_2 V}{2^{s-1}\pi s}\left(\frac{b-a}{\ell' - s}\right)^s.$$

Similarly, we bound the third term as follows

$$\|\|v\|_2 V_\ell p_{\ell'}(T_\ell)e_1 - \|v\|_2 V_\ell f(T_\ell)e_1\|_2 \leq \frac{\|v\|_2 V}{2^{s-1}\pi s}\left(\frac{b-a}{\ell' - s}\right)^s,$$

since V_ℓ is the semi-orthogonal matrix whose 2-norm is 1. To estimate the second term, we use Theorem 8.1 in [60], which relies on the Bernstein ellipse.

Theorem B.2. *[Theorem 8.1 [60]] Let $f(x)$ be analytic in $[-1,1]$ and assume that $f(x)$ can be extended analytically to the open Bernstein ellipse E_ρ for some $\rho > 1$, where it satisfies $|f(x)| \leq M(\rho)$ for some $M(\rho)$. Then, the coefficients of the Chebyshev approximation of the function satisfy $|c_0| \leq M(\rho)$ and*

$$|c_n| \leq 2M(\rho)\rho^{-n}, \quad n \geq 1.$$

Note that the Fermi-Dirac distribution $f(x)$ is analytic in the strip $\{z : |\mathrm{Im}(z)| < \frac{\pi}{\beta}\}$ and $z = \mu \pm \frac{\pi}{\beta} i$ are the singular points. Since the two singular points correspond to $\frac{2}{b-a}\big(\mu - \frac{a+b}{2} \pm \frac{\pi}{\beta} i\big)$ under the linear transformation, the function $\tilde{f}(x)$ is analytic in the scaled strip $\{z : |\mathrm{Im}(z)| < \frac{2}{b-a}\frac{\pi}{\beta}\}$. Thus, by the continuity of the Bernstein

ellipse E_ρ with respect to ρ, we can find ρ sufficiently close to 1 such that a Bernstein ellipse is a proper subset of the strip. Then, we apply the theorem to the function $\tilde{f}(x)$ and consider its Chebyshev approximation $\tilde{p}_{\ell'}(x) = \sum_{n=0}^{\ell'} c_n T_n(x)$. By the scaling in (B.1), we can deduce that

$$p_{\ell'}(A) = \tilde{p}_{\ell'}(\tilde{A}).$$

Thus, we have

$$\|p_{\ell'}(A)v - \|v\|_2 V_\ell p_{\ell'}(T_\ell)e_1\|_2 = \|\sum_{n=\ell+1}^{\ell'} c_n T_n(\tilde{A})v + \|v\|_2 V_\ell \sum_{n=\ell+1}^{\ell'} c_n T_n(\tilde{T}_\ell)e_1\|_2$$

$$\leq 2\sum_{n=\ell+1}^{\ell'} |c_n|\|v\|_2 \leq 2\sum_{n=\ell+1}^{\infty} |c_n|\|v\|_2 = 4M(\rho)\|v\|_2 \frac{\rho^{-\ell}}{\rho-1}.$$

In the first equality, we applied Lemma 3.1 in [54], which states as

$$p_j(A)v = \|v\|_2 V_\ell p_j(T_\ell)e_1$$

for any polynomial $p(x)$ of degree $j \leq \ell$. In addition, the first inequality holds since $|T_n(x)| \leq 1$ and $\|\tilde{T}_\ell\| \leq \|\tilde{A}\| \leq 1$. In the last step, we have used the theorem above.

Now we are ready to prove Theorem 4.3.

Proof. By collecting the above results, the error is bounded by,

$$\|f(A)v - \|v\|_2 V_\ell f(T_\ell)e_1\|_2 \leq \frac{\|v\|_2 V}{2^{s-2}\pi s}\left(\frac{\lambda_{\max}(A) - \lambda_{\min}(A)}{\ell' - s}\right)^s + 4M(\rho)\|v\|_2 \frac{\rho^{-\ell}}{\rho-1}.$$

Recalling that ℓ' is arbitrary and greater than ℓ, the first term on the right hand side can be removed by letting $\ell' \to \infty$. This completes the proof. □

Appendix C. The proof of Theorem 3.3

Proof. By the assumption that $\mathrm{Re}(\lambda) < 1$ for each eigenvalue λ of $K'(q^*)$, there exists a small $\theta_{\max} > 0$ such that for any $0 < \theta < \theta_{\max}$, the spectral radius of $K'_\theta(q^*)$ is less than 1. This is similar to the standard stability condition for the Euler's method for solving ODEs, and it corresponds to a circular disk in the complex plane. We choose such a parameter θ. We observe that $K'_{U,\theta}(p^*)$ in the theorem 3.3 reduces to an upper triangular matrix from the Schur decomposition of $K'(q^*)$ with some unitary matrix U. Theorems 3 and 4 in [30][p 214] provide an explicit construction of a matrix norm of $K'_{U,\theta}(p^*)$ that is less than 1. The norm is induced by an inner product using a diagonal matrix, $D \succ 0$. We denote the vector norm by $\|\cdot\|_D$. Those theorems and the choice of θ yield that, $\forall p_1$ and p_2,

$$\|K'_{U,\theta}(p^*)(p_1 - p_2)\|_D \leq c_1\|p_1 - p_2\|_D, \tag{C.1}$$

for some $c_1 < 1$. Now, we are ready to prove the result. Since K is continuously differentiable, $K_{U,\theta}$ in Theorem 3.3 is continuously differentiable as well. Thus, by the continuity of the Jacobian, we have $\|K'_{U,\theta}(p)\|_D < c_2 < 1$, for some c_2 and for any $p \in B_D(p^*, \rho)$ where $B_D(p^*, \rho)$ is a ball centered at p^* of some radius ρ with respect to the D-norm. The local contractiveness can then be checked using the mean-value theorem for vector-valued functions. Since θ is arbitrarily chosen in $(0, \theta_{\max})$, the proof is completed.

□

Appendix D. Lemmas for the proofs in section 3.1

Lemma D.1. *For each $N \in \mathbb{N}$ and a positive number v, assume that $\mathbb{P}(\|e_n\|_2 > v$ for all $n \geq N) = 0$. Then,*

$$\mathbb{P}(\liminf_n \|e_n\|_2 \leq v) = 1.$$

Proof. Let $A_N := \{w : \|e_n\|_2 > v \text{ for all } n \geq N\}$. Note that $A_N \subset A_{N+1}$. Then,

$$\bigcup_N A_N = \{w : \text{there exists a } N \in \mathbb{N} \text{ such that } \|e_n\|_2 > v \text{ for all } n \geq N\},$$

which implies

$$\left(\bigcup_N A_N\right)^c = \{w : \liminf_n \|e_n\|_2 \leq v\}.$$

By the countable additivity, therefore, the Lemma holds true. □

However, for case $m \geq 2$, we will develop a more sophisticated tool. In the following Lemma, we employ well known results on irreducible aperiodic stochastic matrices in [37, 47]. Moreover, we will use the Perron-Frobenius theorem in [44].

Lemma D.2. *Suppose that a sequence of iterates $\{e_n\}$ from the linear mixing scheme 1 is bounded. Assume that for fixed $b_m > 0$,*

$$\lim_{n\to\infty} \left[\|e_{n+m-1}\|_2^2 + \sum_{j=2}^{m} \left(\sum_{i=1}^{m-j+1} b_i \right) \|e_{n+m-j}\|_2^2 \right] = x. \tag{D.1}$$

Then,

$$\lim_{n\to\infty} \|e_n\|_2 = \sqrt{\frac{x}{1 + \sum_{j=2}^{m} \left(\sum_{i=1}^{m-j+1} b_i \right)}}.$$

Proof. Take a subsequence $\{e_{n_k}\}$. Then, since the sequence $\{e_n\}$ is bounded, we can find a convergent sub-subsequence. To reduce notations, we preserve the same indices $\{n_k\}$ for this convergent sub-subsequence. Furthermore, we can assume that the m shifted sequences are convergent, namely,

$$\{e_{n_k}\}, \{e_{n_k-1}\}, ..., \{e_{n_k-(m-1)}\} \text{ converge.}$$

For any $N \in \mathbb{N}$ and $N > m$, let $l_{i,N} := \lim_{k\to\infty} e_{n_k-N+i}$ for $1 \leq i \leq N$. Then, it follows that for $n \in \{1, 2, ..., N-m\}$,

$$l_{n+m,N} = \sum_{i=1}^{m} b_i l_{n+i-1,N}, \tag{D.2}$$

from the mixing scheme 1 by noting that the damping parameters $\{a_n\}$ converge to 0. We claim that the limits of the shifted sequences are the same, i.e.,

$$\lim_{k\to\infty} e_{n_k} = \lim_{k\to\infty} e_{n_k-1} = \cdots = \lim_{k\to\infty} e_{n_k-(m-1)}.$$

It is sufficient to show that the first entries of the limits are the same. With the standard basis vector $e_1 = (1, 0, 0, .., 0)^T$, we define the nth vector

$$\tilde{l}_{n,N} := (l_{n+m-1,N} \cdot e_1, l_{n+m-2,N} \cdot e_1, ..., l_{n,N} \cdot e_1)^T,$$

which contains the first entries of the vectors $\{l_{i,N}\}_{i=n}^{n+m-1}$. Next, from the relation (D.2), we can define a recursive system such that for $1 \leq n \leq N-m$,

$$\tilde{l}_{n+1,N} = B\tilde{l}_{n,N},$$

where

$$B = \begin{pmatrix} b_m & b_{m-1} & \cdots & b_2 & b_1 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & 1 & 0 \end{pmatrix}.$$

The matrix B can be recognized as a companion matrix. Thus, the characteristic polynomial of B has one as its root, because $\sum_{i=1}^{m} b_i = 1$.

Note that it is the mixing scheme with m steps, which assumes that $b_1 > 0$. For this reason, B is an irreducible matrix, which means that all the nodes $\{1, 2, .., m\}$ communicate in the graph corresponding to the matrix B [37][p 86], which can interpreted as a transition matrix.

Moreover, since $b_m > 0$ by assumption and B is irreducible, B is aperiodic [37][p 91]. Also, since all rows sum to one, B is a stochastic matrix. By the Gershgorin's theorem, we can guarantee that $\rho(B) \leq 1$, which denotes the spectral radius of B. Thus, by applying the Perron-Frobenius theorem to B^T [44][p 673], we can find a left eigenvector $\pi > 0$

$$\pi B = \pi.$$

Since B is irreducible, aperiodic and has the invariant distribution π, the matrix B^n converges to equilibrium as stated in [47][Theorem 1.8.3], namely,

$$\lim_{n\to\infty} B^n = \begin{pmatrix} \pi_1 & \pi_2 & \cdots & \pi_m \\ \pi_1 & \pi_2 & \cdots & \pi_m \\ \pi_1 & \pi_2 & \cdots & \pi_m \\ \vdots & \vdots & \cdots & \vdots \\ \pi_1 & \pi_2 & \cdots & \pi_m \end{pmatrix}.$$

On the other hand, from the recursive relation, we have $\tilde{l}_{N-m+1,N} = B^{N-m}\tilde{l}_{1,N}$, or

$$\begin{pmatrix} l_{N,N} \cdot e_1 \\ l_{N-1,N} \cdot e_1 \\ \vdots \\ l_{N-m+1,N} \cdot e_1 \end{pmatrix} = B^{N-m} \begin{pmatrix} l_{m,N} \cdot e_1 \\ l_{m-1,N} \cdot e_1 \\ \vdots \\ l_{1,N} \cdot e_1 \end{pmatrix}.$$

Since the sequence $\{e_n\}_{n=1}^{\infty}$ is bounded by assumption and B^n converges, by letting $N \to \infty$, we can obtain the result that for any $i, j \in \{0, 1, 2, ..., m-1\}$,

$$\lim_{N\to\infty} l_{N-i,N} \cdot e_1 = \lim_{N\to\infty} l_{N-j,N} \cdot e_1,$$

which implies that

$$\lim_{k\to\infty} e_{n_k} \cdot e_1 = \lim_{k\to\infty} e_{n_k-1} \cdot e_1 = \cdots = \lim_{k\to\infty} e_{n_k-m+1} \cdot e_1.$$

With the same technique for the other coordinates, it immediately follows that

$$\lim_{k\to\infty} e_{n_k} = \lim_{k\to\infty} e_{n_k-1} = \cdots = \lim_{k\to\infty} e_{n_k-m+1}.$$

Here, we proved the claim. From this result, we can use Assumption (D.1) as follows

$$x = \lim_{k\to\infty}\left[\|e_{n_k}\|_2^2 + \sum_{j=2}^{m}\left(\sum_{i=1}^{m-j+1} b_i\right)\|e_{n_k+1-j}\|_2^2\right] = \left[1+\sum_{j=2}^{m}\left(\sum_{i=1}^{m-j+1} b_i\right)\right]\| \lim_{k\to\infty} e_{n_k}\|_2^2.$$

To sum up, for any subsequence of the sequence $\{\|e_n\|_2\}_{n=1}^{\infty}$, we can find a sub-subsequence convergent to

$$\sqrt{\frac{x}{1+\sum_{j=2}^{m}\left(\sum_{i=1}^{m-j+1} b_i\right)}}.$$

Therefore, this completes the proof of the lemma.

□

Next, in order to derive non-negative supermartingales from the general mixing scheme 1 with the Lyapunov functions (3.11), we prove the following inequalities which will be used in Theorem 3.2.

By using the Cauchy-Schwarz inequality and the Jensen's inequality, the $n+1$st error can be bounded as

$$\begin{aligned}
&\mathbb{E}[\|e_{n+1}\|_2^2] = \mathbb{E}[\|B_m(e_n) + a_n B_m(G_n)\|_2^2] \\
&\leq \sum_{i=1}^{m} b_i^2 \mathbb{E}[\|e_i + a_n G_i\|_2^2] + \sum_{i\neq j}\mathbb{E}[b_i b_j (e_i + a_n G_i, e_j + a_n G_j)] \\
&\leq \sum_{i=1}^{m} b_i^2 \mathbb{E}[\|e_i + a_n G_i\|_2^2] + \sum_{i\neq j}\sqrt{b_i^2\mathbb{E}[\|e_i + a_n G_i\|_2^2]}\sqrt{b_j^2\mathbb{E}[\|e_j + a_n G_j\|_2^2]} \\
&= \left(\sum_{i=1}^{m} b_i\sqrt{\mathbb{E}[\|e_i + a_n G_i\|_2^2]}\right)^2 \leq \sum_{i=1}^{m} b_i \mathbb{E}[\|e_i + a_n G_i\|_2^2]
\end{aligned} \tag{D.3}$$

With this, we can establish inequalities which will define a non-negative supermartingale eventually.

For general m, from the inequality (D.3), the conditional expectation of $\|X_{n+1}\|_{n+1}$ is bounded as

$$\begin{aligned}
&\mathbb{E}[\|X_{n+1}\|_{n+1}|X_n] \\
&= \mathbb{E}[\|e_{n+m}\|_2^2|X_n] + \sum_{j=2}^{m}\sum_{i=j}^{m} b_{m-i+1}\mathbb{E}[\|e_{n+j-1+m-i} + a_{n+m-2+j}G_{n+j-1+m-i}\|_2^2|X_n] \\
&\leq \sum_{i=1}^{m} b_i \mathbb{E}[\|e_{n+i-1} + a_{n+m-1}G_{n+i-1}\|_2^2|X_n] \\
&+ \sum_{j=2}^{m}\sum_{i=j}^{m} b_{m-i+1}\mathbb{E}[\|e_{n+j-1+m-i} + a_{n+m-2+j}G_{n+j-1+m-i}\|_2^2|X_n].
\end{aligned}$$

In this upper bound, the first summation is divided into two parts as follows

$$b_m\mathbb{E}[\|e_{n+m-1} + a_{n+m-1}G_{n+m-1}\|_2^2|X_n] + \sum_{i=1}^{m-1} b_i\|e_{n+i-1} + a_{n+m-1}G_{n+i-1}\|_2^2.$$

Meanwhile, the double summation term equals

$$\sum_{j=2}^{m} b_{m-j+1}\mathbb{E}[\|e_{n+m-1}+a_{n+m-2+j}G_{n+m-1}\|_2^2|X_n] \\ +\sum_{j=2}^{m-1}\sum_{i=j+1}^{m} b_{m-i+1}\|e_{n+j-1+m-i}+a_{n+m-2+j}G_{n+j-1+m-i}\|_2^2,$$

by pulling out the terms involving e_{n+m-1}. By grouping terms with and without e_{n+m-1}, we can rewrite the bound as

$$\mathbb{E}[\|X_{n+1}\|_{n+1}|X_n] \le \sum_{j=1}^{m} b_{m-j+1}\mathbb{E}[\|e_{n+m-1}+a_{n+m-2+j}G_{n+m-1}\|_2^2|X_n] \\ +\sum_{j=2}^{m}\sum_{i=j}^{m} b_{m-i+1}\|e_{n+j-2+m-i}+a_{n+m-3+j}G_{n+j-2+m-i}\|_2^2.$$

By recalling Definition (D.7), it follows that

$$\begin{aligned} &\mathbb{E}[\|X_{n+1}\|_{n+1}|X_n]-\|X_n\|_n \\ &=\sum_{j=1}^{m} b_{m-j+1}\mathbb{E}[\|e_{n+m-1}+a_{n+m-2+j}G_{n+m-1}\|_2^2|X_n]-\|e_{n+m-1}\|_2^2 \\ &\le 2\Big(\sum_{j=1}^{m} b_{m-j+1}a_{n+m-2+j}\Big)(e_{n+m-1},R_{n+m-1}) \\ &+\sum_{j=1}^{m} b_{m-j+1}a_{n+m-2+j}^2\Big(\Xi+\|R_{n+m-1}\|^2\Big) \\ &\le -(1-c)\Big(\sum_{j=1}^{m} b_{m-j+1}a_{n+m-2+j}\Big)\|e_{n+m-1}\|_2^2 \\ &+\Xi\sum_{j=1}^{m} b_{m-j+1}a_{n+m-2+j}^2. \end{aligned} \tag{D.4}$$

In the first inequality, we used Assumption 2 and the fact that $(R_{n+m-1},\xi_{n+m-1})=0$ together with Definition (3.4). In the second inequality, we apply the inequalities (E.1) and the condition of boundedness in (3.7). From the last part of the inequality D.4, we make notation for the coefficients attached to the terms $\|e_{n+m-1}\|^2$ and Ξ as follow

$$A_n := \textstyle\sum_{j=1}^{m} b_{m-j+1}a_{n+m-2+j} \tag{D.5a}$$

$$\chi_n := \textstyle\sum_{j=1}^{m} b_{m-j+1}a_{n+m-2+j}^2. \tag{D.5b}$$

Note that the condition (3.7) together with (3.9) yields

$$\sum_{n=1}^{\infty} A_n=\infty, \quad \sum_{n=1}^{\infty}\chi_n<\infty. \tag{D.6}$$

By defining $V_n(X_n)=\|X_n\|_n+\Xi\sum_{i=n}^{\infty}B_i$, the above can be rewritten

$$\mathbb{E}[V_{n+1}(X_{n+1})|X_n]-V_n(X_n)\le -(1-c)A_n\|e_{n+m-1}\|_2^2\le 0. \tag{D.7}$$

Appendix E. The proofs of theorems in section 4

E.1. **Theorem 3.1.**

Proof. To establish the stability, we follow the proof in [36][p 112, Theorem 5.1]. By the definition of $V(\cdot)$, for any $q_n \in B(q^*, \rho)$ direct calculations yield

$$\begin{aligned}
&\mathbb{E}_n\left[V(q_{n+1})\right] - V(q_n) = \mathbb{E}_n\left[\|e_n + a_n G_n\|_2^2\right] - \|e_n\|_2^2, \\
&= 2a_n \mathbb{E}_n\left[(e_n, G_n)\right] + a_n^2 \mathbb{E}_n\left[\|G_n\|_2^2\right], \\
&\leq -2a_n(1-c)\|e_n\|_2^2 + (c+1)^2 a_n^2 \|e_n\|_2^2 + \Xi a_n^2, \\
&\leq -a_n(1-c)\|e_n\|_2^2 + \Xi a_n^2.
\end{aligned}$$

In the first inequality, we use Assumption 2 and remove the cross term $\mathbb{E}_n[(R_n, \xi_n)]$, where the residual R_n and the error ξ_n sum to G_n. Besides, we used the following inequalities

$$(e_n, R_n) = -\|e_n\|_2^2 + (K(q_n) - K(q^*), e_n) \leq (c-1)\|e_n\|_2^2 \tag{E.1a}$$

$$\|R_n\|^2 \leq (1+c)^2 \|e_n\|_2^2. \tag{E.1b}$$

In the last step, we used the condition that $a_n \leq \frac{1-c}{(1+c)^2}$.

We proceed by observing that $V_n(q_n) \geq 0$ and

$$\delta V_{n+1} - \delta V_n = -\Xi a_n^2,$$

which implies the following inequality,

$$\mathbb{E}_n[V_{n+1}(q_{n+1})] - V_n(q_n) \leq -(1-c)a_n\|e_n\|_2^2 \leq 0. \tag{E.2}$$

Here, we define the stopping time $\tau_\rho := \{n : \|e_n\|_2 > \rho\}$. Accordingly, we define a stopped process for q_n and a corresponding Lyapunov function as follows

$$\tilde{q}_n := \begin{cases} q_n, \ n \leq \tau_\rho \\ q_{\tau_\rho}, \ n > \tau_\rho \end{cases}, \tilde{V}_n := \begin{cases} V_n, \ n \leq \tau_\rho \\ V_{\tau_\rho}, \ n > \tau_\rho \end{cases}. \tag{E.3}$$

This technique is shown in the proof of Theorem 5.1 in [36], which yields the non-negative supermartingale $\{\tilde{V}_n(\tilde{q}_n)\}$.

Similar to the proof of Theorem 5.1 in [36], we can deduce that,

$$\mathbb{P}\left\{\sup_n \|e_n\|_2 > \rho | q_1\right\} \mathbb{I}_{\{q_1 \in B(q^*,\rho)\}} \leq \mathbb{P}\left\{\sup_n V_n(q_n) > \rho^2 | q_1\right\} \mathbb{I}_{\{q_1 \in B(q^*,\rho)\}} \leq \frac{V_1(q_1)}{\rho^2},$$

which concludes the first part of the theorem.

Since the stopped process $\{\tilde{V}_n(\tilde{q}_n)\}_{n\geq 1}$ forms a supermartingale as

$$\mathbb{E}_n[\tilde{V}_{n+1}(\tilde{q}_{n+1})] \leq \tilde{V}_n(\tilde{q}_n), \ \forall n \geq 1, \tag{E.4}$$

$\tilde{V}_n(\tilde{q}_n)$ converges to some random variable $\tilde{V} \geq 0$. In the event where $\|e_n\|_2 \leq \rho$ for all $n \in \mathbb{N}$, this implies that

$$\lim_{n\to\infty} V_n(q_n) = \lim_{n\to\infty} \|e_n\|_2^2,$$

with probability one, since $\sum_n a_n^2 < \infty$. Suppose that $\|e_n\|_2$ converges to a positive random variable V with positive probability. Then, there exists a positive number $\delta > 0$ such that

$$\mathbb{P}\left\{\lim_{n\to\infty} \|e_n\|_2 > \delta \middle| \{q_n\} \subset B(q^*, \rho)\right\} > 0.$$

By Lemma D.1, we have for some $N \in \mathbb{N}$,

$$\mathbb{P}\left\{\|e_n\|_2 > \frac{\delta}{2} \text{ for all } n \geq N \middle| \lim_{n\to\infty} \|e_n\| > \delta, \{q_n\} \subset B(q^*, \rho)\right\} > 0.$$

On the other hand, by a telescoping trick with the inequality (E.2), for any given $q_1 \in B(q^*, \rho)$, we have

$$V_1(q_1) \geq V_1(q_1) - \mathbb{E}_1[\tilde{V}_n(\tilde{q}_n)] \geq 2(1-c)\mathbb{E}_1\left[\sum_{i=1}^{n-1} a_i \|\tilde{e}_i\|_2^2\right],$$

which implies

$$\mathbb{E}_1\left[\sum_{i=1}^{\infty} a_i \|\tilde{e}_i\|_2^2\right] < \infty.$$

By the above results, we can deduce that

$$\mathbb{P}\left\{\|e_n\|_2 \geq \frac{\delta}{2} \text{ for all } n \geq N, \{q_n\}_{n=1}^{\infty} \subset B(q^*, \rho)\right\} > 0,$$

which implies that

$$\begin{aligned}\infty > \mathbb{E}_1\left[\sum_{i=1}^{\infty} a_i \|\tilde{e}_i\|_2^2\right] &\geq \mathbb{E}_1\left[\sum_{i=1}^{\infty} a_i \|\tilde{e}_i\|_2^2 \mathbb{I}_{\{\|e_n\|_2 > \frac{\delta}{2} \text{ for all } n \geq N, \{q_n\}_{n=1}^{\infty} \subset B(q^*,\rho)\}}\right] \\ &\geq \frac{\delta^2}{4}\left(\sum_{i=N}^{\infty} a_i\right) \cdot \mathbb{P}\left\{\|e_n\|_2 > \frac{\delta}{2} \text{ for all } n \geq N, \{q_n\}_{n=1}^{\infty} \subset B(q^*, \rho)\right\}.\end{aligned}$$

Since $\sum_n a_n = \infty$, this is a contradiction. Therefore, $\|e_n\|_2$ converges to 0 with probability one when $\{q_n\} \subset B(q^*, \rho)$. □

E.2. **Theorem 3.2.**

Proof. Define $\tau_\rho = \min\{n : \|e_{n+m-1}\|_2 > \rho\}$ as a stopping time. Similar to (E.3), we define

$$\tilde{X}_n := \begin{cases} X_n, \ n \leq \tau_\rho \\ X_{\tau_\rho}, \ n > \tau_\rho \end{cases}, \tilde{V}_n := \begin{cases} V_n, \ n \leq \tau_\rho \\ V_{\tau_\rho}, \ n > \tau_\rho. \end{cases} \tag{E.5}$$

We will prove the theorem similar to the proof of the theorem 3.1. The inequality (D.7) will yield a non-negative supermartingale, which justifies the first statement. Also, the stopped process $\tilde{V}_n(\tilde{X}_n)$ converges to some random variable $\tilde{V} \geq 0$. Therefore, in the event where $\|e_n\|_2 \leq \rho$ for all $n \in \mathbb{N}$, we can deduce that

$$\begin{aligned}\lim_{n\to\infty} V_n(X_n) = \lim_{n\to\infty} \|X_n\|_n &= \lim_{n\to\infty}\left[\|e_{n+m-1}\|_2^2 + \sum_{j=2}^{m}\sum_{i=j}^{m} b_{m-i+1}\|e_{n+j-2+m-i}\|_2^2\right] \\ &= \lim_{n\to\infty}\left[\|e_{n+m-1}\|_2^2 + \sum_{j=2}^{m}\left(\sum_{i=1}^{m-j+1} b_i\right)\|e_{n+m-j}\|_2^2\right]\end{aligned}$$

with probability one. The second equality holds as in the previous proof. Moreover, by applying the lemma D.2 to the last step, the sequence $\{\|e_n\|_2\}$ converges with

probability one, namely,

$$\lim_{n\to\infty} V_n(X_n) = \left[1 + \sum_{j=2}^{m}\left(\sum_{i=1}^{m-j+1} b_i\right)\right] \lim_{n\to\infty} \|e_n\|_2^2$$

For similar reasoning in the proof of the theorem 3.1, $\|e_n\|_2$ converges to 0 when all the iterates are in $B(q^*, \rho)$.

□

E.3. **Theorem 3.5.**

Proof. As already established, we use the almost supermartingale property (D.7). This property can be rewritten as

$$\mathbb{E}[V_{n+1}(X_{n+1})|X_n] \leq V_n(X_n) - A_n k(X_n), \tag{E.6}$$

where A_n is defined in (D.5) and $k(X_n) = (1-c)\|e_{n+m-1}\|_2^2$. With (E.5) and this function $k(X_n)$, we can define the non-negative supermartingale as similar in Theorem 5.1 [36]

$$\mathbb{E}[\tilde{V}_{n+1}(\tilde{X}_{n+1})|\tilde{X}_n] \leq \tilde{V}_n(\tilde{X}_n) - A_n \tilde{k}(\tilde{X}_n), \tag{E.7}$$

where

$$\tilde{k}(\tilde{X}_n) = \begin{cases} k(X_n),\ q_{n+m-1} \in B(q^*, \rho) \\ 0,\ q_{n+m-1} \notin B(q^*, \rho). \end{cases} \tag{E.8}$$

By taking the total expectation on the supermartingale and telescoping inequalities, we obtain

$$\sum_{n=1}^{j} A_n \mathbb{E}[k(X_n)\mathbb{I}_{E_j}|X_1] \leq \sum_{n=1}^{j} A_n \mathbb{E}[\tilde{k}(\tilde{X}_n)|X_1] \leq V_1(X_1). \tag{E.9}$$

We note that as defined above, the function $k(X_n)$ is a convex function with respect to q_{n+m-1} as a quadratic function. Thus, by the Jensen's inequality, we have

$$(1-c)\|\bar{q} - q^*\|_2^2 \leq \sum_{n=1}^{j}\left(\frac{A_n}{\sum_{n=1}^{j} A_n}\right) k(X_n), \tag{E.10}$$

where

$$\bar{q} := \sum_{n=1}^{j}\left(\frac{A_n}{\sum_{n=1}^{j} A_n}\right) q_n. \tag{E.11}$$

To put these together, we arrive at

$$\mathbb{E}[\|\bar{q} - q^*\|_2^2 \mathbb{I}_{E_j}|X_1] \leq \frac{V_1(X_1)}{(1-c)\left(\sum_{n=1}^{j} A_n\right)}. \tag{E.12}$$

This is the first result. For convenience, let us denote the RHS of this inequality by u_j.

By applying the Markov's inequality to this result, we have

$$\mathbb{P}\left\{\|\bar{q} - q^*\|_2 > \epsilon \text{ and } E_j \text{ occurs } |X_1\right\} \leq \frac{u_j}{\epsilon^2},$$

equivalently,

$$\mathbb{P}\left\{\|\bar{q} - q^*\|_2 \leq \epsilon \text{ or } E_j \text{ does not occur}|X_1\right\} \geq 1 - \frac{u_j}{\epsilon^2}.$$

By the stability result in the theorem 3.2, the probability for E_j to not occur is bounded above

$$\mathbb{P}\left\{E_j \text{ does not occur}|X_1\right\} \leq \frac{V_1(X_1)}{\rho^2},$$

which leads to

$$\mathbb{P}\left\{\|\bar{q} - q^*\|_2 \leq \epsilon|X_1\right\} \geq 1 - \frac{u_j}{\epsilon^2} - \frac{V_1(X_1)}{\rho^2}.$$

□

References

[1] Y. I. Alber, C. Chidume, and J. Li, *Stochastic approximation method for fixed point problems*, Applied Mathematics **3** (2012), no. 12, 2123–2132.

[2] H. Avron and S. Toledo, *Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix*, Journal of the ACM **58** (April 2011), no. 2, 1–34 (en).

[3] A. S. Banerjee, P. Suryanarayana, and J. E. Pask, *Periodic pulay method for robust and efficient convergence acceleration of self-consistent field iterations*, Chemical Physics Letters **647** (2016), 31–35.

[4] T. L. Beck, *Real-space mesh techniques in density-functional theory*, Reviews of Modern Physics **72** (October 2000), no. 4, 1041–1080 (en).

[5] C. Bekas, E. Kokiopoulou, and Y. Saad, *An estimator for the diagonal of a matrix*, Applied Numerical Mathematics **57** (2007), no. 11-12, 1214–1229.

[6] L. Bottou and O. Bousquet, *The tradeoffs of large scale learning*, Advances in neural information processing systems **20** (2007).

[7] L. Bottou, F. E. Curtis, and J. Nocedal, *Optimization methods for large-scale machine learning*, SIAM Review **60** (2018), no. 2, 223–311.

[8] D. R. Bowler and M. J. Gillan, *An efficient and robust technique for achieving self consistency in electronic structure calculations*, Chemical Physics Letters **325** (2000), no. 4, 473–476.

[9] D. R. Bowler, T. Miyazaki, and M. J. Gillan, *Recent progress in linear scaling ab initio electronic structure techniques*, Journal of Physics: Condensed Matter **14** (2002), no. 11, 2781.

[10] E. Cancès, *Self-consistent field algorithms for Kohn–Sham models with fractional occupation numbers*, The Journal of Chemical Physics **114** (2001), no. 24, 10616–10622.

[11] E. Cancès, G. Kemlin, and A. Levitt, *Convergence analysis of direct minimization and self-consistent iterations*, SIAM Journal on Matrix Analysis and Applications **42** (2021), no. 1, 243–274.

[12] E. Cancès and C. Le Bris, *Can we outperform the DIIS approach for electronic structure calculations?*, International Journal of Quantum Chemistry **79** (2000), no. 2, 82–90.

[13] ———, *On the convergence of SCF algorithms for the hartree-fock equations*, ESAIM: Mathematical Modelling and Numerical Analysis **34** (2000), no. 4, 749–774.

[14] K. L. Chung, *On a stochastic approximation method*, The Annals of Mathematical Statistics (1954), 463–483.

[15] Y. Cytter, E. Rabani, D. Neuhauser, and R. Baer, *Stochastic density functional theory at finite temperatures*, Physical Review B **97** (2018), no. 11, 115207.

[16] A. Defazio, F. Bach, and S. Lacoste-Julien, *Saga: A fast incremental gradient method with support for non-strongly convex composite objectives*, arXiv preprint arXiv:1407.0202 (2014).

[17] F. Diele, I. Moret, and S. Ragni, *Error estimates for polynomial Krylov approximations to matrix functions*, SIAM journal on matrix analysis and applications **30** (2009), no. 4, 1546–1565.

[18] M. Eiermann and O. G. Ernst, *A restarted Krylov subspace method for the evaluation of matrix functions*, SIAM Journal on Numerical Analysis **44** (2006), no. 6, 2481–2504.

[19] M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, Th. Frauenheim, S. Suhai, and G. Seifert, *Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties*, Physical Review B **58** (1998), no. 11, 7260.

[20] H. Fang and Y. Saad, *Two classes of multisecant methods for nonlinear acceleration*, Numerical Linear Algebra with Applications **16** (2009), no. 3, 197–221.

[21] C. J. García-Cervera, J. Lu, and W. E, *A sub-linear scaling algorithm for computing the electronic structure of materials*, Communications in Mathematical Sciences **5** (2007), no. 4, 999–1026.
[22] S. Goedecker, *Linear scaling electronic structure methods*, Reviews of Modern Physics **71** (1999), no. 4, 1085.
[23] F. Golse, S. Jin, and T. Paul, *The random batch method for N-Body quantum dynamics*, arXiv preprint arXiv:1912.07424 (2019).
[24] T. P. Hamilton and P. Pulay, *Direct inversion in the iterative subspace (DIIS) optimization of open-shell, excited-state, and small multiconfiguration scf wave functions*, The Journal of Chemical Physics **84** (1986), no. 10, 5728–5734.
[25] J. Hermann, Z. Schätzle, and F. Noé, *Deep-neural-network solution of the electronic Schrödinger equation*, Nature Chemistry **12** (2020), no. 10, 891–897.
[26] P. Hohenberg and W. Kohn, *Inhomogeneous electron gas*, Physical Review **136** (1964), no. 3B, B864.
[27] S. Jin and X. Li, *Random batch algorithms for quantum Monte Carlo simulations*, Communications in Computational Physics **28** (2020), no. 5, 1907–1936.
[28] D. D. Johnson, *Modified Broyden's method for accelerating convergence in self-consistent calculations*, Physical Review B **38** (1988), no. 18, 12807.
[29] R. Johnson and T. Zhang, *Accelerating stochastic gradient descent using predictive variance reduction*, Advances in Neural Information Processing Systems **26** (2013), 315–323.
[30] D. Kincaid, D. R. Kincaid, and E. W. Cheney, *Numerical analysis: mathematics of scientific computing*, Vol. 2, American Mathematical Soc., 2009.
[31] W. Kohn and L. J. Sham, *Self-consistent equations including exchange and correlation effects*, Physical Review **140** (1965), no. 4A, A1133–A1138.
[32] G. Kresse and J. Furthmüller, *Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set*, Physical Review B **54** (1996), no. 16, 11169.
[33] L. Kronik, A. Makmal, M. L. Tiago, M. Alemany, M. Jain, X. Huang, Y. Saad, and J. R. Chelikowsky, *PARSEC–the pseudopotential algorithm for real-space electronic structure calculations: recent advances and novel applications to nano-structures*, Physica status solidi (b) **243** (2006), no. 5, 1063–1079.
[34] H. J. Kushner, *On the stability of stochastic dynamical systems*, Proceedings of the National Academy of Sciences of the United States of America **53** (1965), no. 1, 8.
[35] ———, *Stochastic stability and control*, Academic Press, New York, 1967.
[36] H. J. Kushner and G. G. Yin, *Stochastic approximation and recursive algorithms and applications*, Vol. 35, Springer Science & Business Media, 2003.
[37] M. Lefebvre, *Applied stochastic processes*, Springer Science & Business Media, 2007.
[38] L. Lin, Y. Saad, and C. Yang, *Approximating spectral densities of large matrices*, SIAM review **58** (2016), no. 1, 34–65.
[39] L. Lin and C. Yang, *Elliptic preconditioner for accelerating the self-consistent field iteration in kohn–sham density functional theory*, SIAM Journal on Scientific Computing **35** (2013), no. 5, S277–S298.
[40] M. A. Marques, A. Castro, G. F. Bertsch, and A. Rubio, *octopus: a first-principles tool for excited electron–ion dynamics*, Computer Physics Communications **151** (2003), no. 1, 60–78.
[41] R. M. Martin, *Electronic Structure: Basic Theory and Practical Methods*, Cambridge University Press, 2011.
[42] PG Martinsson and J. Tropp, *Randomized numerical linear algebra: Foundations & algorithms*, arXiv preprint arXiv:2002.01387 (2020).
[43] D. Marx and J. Hutter, *Ab initio molecular dynamics: basic theory and advanced methods*, Cambridge University Press, 2009.
[44] C. D. Meyer, *Matrix analysis and applied linear algebra*, Vol. 71, SIAM, 2000.
[45] M. A. Morales-Silva, K. D. Jordan, L. Shulenburger, and L. K. Wagner, *Frontiers of stochastic electronic structure calculations*, AIP Publishing LLC, 2021.
[46] L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč, *Sarah: A novel method for machine learning problems using stochastic recursive gradient*, International conference on machine learning, 2017, pp. 2613–2621.
[47] J. R. Norris, *Markov chains*, Cambridge University Press, 1998.
[48] R. G. Parr and W. Yang, *Density-functional theory of atoms and molecules*, Oxford University Press, 1995.

[49] 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*, Reviews of modern physics **64** (1992), no. 4, 1045.
[50] S. J. Reddi, A. Hefny, S. Sra, B. Poczos, and A. Smola, *Stochastic variance reduction for nonconvex optimization*, International conference on machine learning, 2016, pp. 314–323.
[51] S. Resnick, *A probability path*, Springer, 2019.
[52] P. J. Reynolds, D. M. Ceperley, B. J. Alder, and W. A. Lester Jr, *Fixed-node quantum Monte Carlo for molecules*, The Journal of Chemical Physics **77** (1982), no. 11, 5593–5603.
[53] H. Robbins and S. Monro, *A stochastic approximation method*, The Annals of Mathematical Statistics (1951), 400–407.
[54] Y. Saad, *Analysis of some Krylov subspace approximations to the matrix exponential operator*, SIAM Journal on Numerical Analysis **29** (1992), no. 1, 209–228.
[55] J. M. Soler, E. Artacho, J. D. Gale, A. García, J. Junquera, P. Ordejón, and D. Sánchez-Portal, *The SIESTA method for ab initio order-N materials simulation*, Journal of Physics: Condensed Matter **14** (2002), no. 11, 2745.
[56] P. Suryanarayana, V. Gavini, T. Blesgen, K. Bhattacharya, and M. Ortiz, *Non-periodic finite-element formulation of Kohn–Sham density functional theory*, Journal of the Mechanics and Physics of Solids **58** (2010), no. 2, 256–280.
[57] K. Thicke, *Accelerating the computation of density functional theory's correlation energy under random phase approximations*, Ph.D. Thesis, 2019.
[58] A. Toth, J. A. Ellis, T. Evans, S. Hamilton, C. T. Kelley, R. Pawlowski, and S. Slattery, *Local Improvement Results for Anderson Acceleration with Inaccurate Function Evaluations*, SIAM Journal on Scientific Computing **39** (January 2017), no. 5, S47–S65 (en).
[59] A. Toth and C. T. Kelley, *Convergence Analysis for Anderson Acceleration*, SIAM Journal on Numerical Analysis **53** (January 2015), no. 2, 805–819 (en).
[60] L. N. Trefethen, *Approximation theory and approximation practice, extended edition*, SIAM, 2019.
[61] M. E. Tuckerman, *Ab initio molecular dynamics: basic concepts, current trends and novel applications*, Journal of Physics: Condensed Matter **14** (2002), no. 50, R1297.
[62] E Weinan and Jianfeng Lu, *Electronic structure of smoothly deformed crystals: Cauchy-born rule for the nonlinear tight-binding model*, Comm. Pure Appl. Math **63** (2010), no. 11, 1432–1468.
[63] D. Williams, *Probability with martingales*, Cambridge university press, 1991.
[64] J. Wolfowitz, *On the stochastic approximation method of Robbins and Monro*, The Annals of Mathematical Statistics **23** (1952), no. 3, 457–461.
[65] Y. Xi, R. Li, and Y. Saad, *Fast computation of spectral densities for generalized eigenvalue problems*, SIAM Journal on Scientific Computing **40** (2018), no. 4, A2749–A2773.
[66] C. Yang, J. C. Meza, and LW Wang, *A constrained optimization algorithm for total energy minimization in electronic structure calculations*, Journal of Computational Physics **217** (2006), no. 2, 709–721.
[67] X. Zhang, J. Zhu, Z. Wen, and A. Zhou, *Gradient type optimization methods for electronic structure calculations*, SIAM Journal on Scientific Computing **36** (2014), no. 3, C265–C289.
[68] Y. Zhou, Y. Saad, M. L. Tiago, and J. R. Chelikowsky, *Self-consistent-field calculations using chebyshev-filtered subspace iteration*, Journal of Computational Physics **219** (2006), no. 1, 172–184.

Department of Mathematics, The Pennsylvania State University, University Park, PA 16802, USA,
Email address: tuk351@psu.edu

Department of Mathematics, The Pennsylvania State University, University Park, PA 16802, USA,
Email address: Xiantao.Li@psu.edu