SIAM J. SCI. COMPUT.
Vol. 42, No. 1, pp. B1–B29

THE FULL CONFIGURATION INTERACTION QUANTUM MONTE CARLO METHOD THROUGH THE LENS OF INEXACT POWER ITERATION[*]

JIANFENG LU[†] AND ZHE WANG[‡]

Abstract. In this paper, we propose a general analysis framework for inexact power iteration, which can be used to efficiently solve high-dimensional eigenvalue problems arising from quantum many-body problems. Under this framework, we establish the convergence theorems for several recently proposed randomized algorithms, including full configuration interaction quantum Monte Carlo and fast randomized iteration. The analysis is consistent with numerical experiments for physical systems such as the Hubbard model and small chemical molecules. We also compare the algorithms both in convergence analysis and numerical results.

Key words. inexact power iteration, FCIQMC, fast randomized iteration

AMS subject classifications. 65F15, 81-08

DOI. 10.1137/18M1166626

1. Introduction.

In recent years, following the work of full configuration interaction quantum Monte Carlo (FCIQMC) [5, 7], the idea of using randomized or truncated power methods to solve the full configuration interaction (FCI) eigenvalue problem has become quite popular in quantum chemistry literature. From a mathematical point of view, the FCI calculation essentially asks for the smallest eigenvalue of a real symmetric matrix (for ground state calculation) or a few low-lying eigenvalues (for low-lying excited state calculation). The computational challenge lies in the fact that the size of the matrix grows exponentially fast with respect to the number of orbitals/electrons in the chemical system, and thus a brute-force numerical diagonalization method (such as the power method or Lanczos method) does not work except for very small molecules.

The goal of this work is twofold: On the one hand, we want to establish a general framework to understand these recently proposed randomized algorithms. As we shall see, from the viewpoint of numerical linear algebra, these recent methods can be understood as generalizations of the conventional power method when an inexact matrix-vector product is used. As a result, the convergence of these methods can be dealt with by a simple extension of the usual proof of convergence of power methods. A natural consequence of this understanding is that, to compare the various approaches, the crucial part is to determine the error caused by different strategies of inexact matrix-vector multiplication. Using these insights, we will compare a few of the recently proposed randomized or truncated FCI methods analytically, as well as numerically, using the Hubbard model and some small chemical molecules as toy examples.

[*] Submitted to the journal's Computational Methods in Science and Engineering section January 22, 2018; accepted for publication (in revised form) October 15, 2018; published electronically January 2, 2020.
https://doi.org/10.1137/18M1166626
Funding: This work was supported in part by the National Science Foundation under award DMS-1454939.

[†] Departments of Mathematics, Physics, and Chemistry, Duke University, Durham, NC 27708 (jianfeng@math.duke.edu).
[‡] Department of Mathematics, Duke University, Durham, NC 27708 (zhe.wang3@duke.edu).

While the motivation of the study is from FCI calculation in quantum chemistry, these methods can be understood on the general setting of numerical linear algebra, and hence, except in the numerical section, we will not restrict ourselves to the FCI Hamiltonian. For a given real symmetric positive definite matrix $A \in \mathbb{R}^{N\times N}$, we are interested in numerically obtaining the largest eigenvalue and corresponding eigenvector. It is possible to extend the method to leading k eigenvalues, where k is on the order of 1 based on the subspace iteration method, a generalization of the power method. In what follows, we denote the eigenvalues of A by $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0$, and corresponding orthonormal eigenvectors are $u_1, u_2, \ldots, u_N$ (viewed as column vectors).

To obtain the largest eigenvalue and the corresponding eigenvector, one of the simplest algorithms is the standard power iteration, given by

$$\begin{aligned} y_{t+1} &= Ax_t, \\ x_{t+1} &= y_{t+1}/\|y_{t+1}\|_2, \end{aligned}$$

with some initial guess x_0 and iterate till convergence. The algorithm is simple to understand: The matrix multiplication Ax_t amplifies x_t in the leading eigenspace. The convergence of the algorithm is also well known: As long as the initial vector satisfies that $u_1^\top x_0$ is nonzero and there exists an eigengap ($\lambda_1 > \lambda_2$), the subspace $\operatorname{span} x_t$ converges to the eigenspace $\operatorname{span} u_1$ linearly as $t \to \infty$ with rate proportional to λ_2/λ_1.

Since only the convergence of the subspace is of interest, the norm of the vector x_t plays no role. Hence the normalization step of power iteration may be omitted:

$$v_{t+1} = Av_t.$$

This is equivalent to the original power method. Of course, in practical computations, the normalization is important to avoid issues like arithmetic overflow.

Motivated by recently proposed algorithms in the quantum chemistry literature, in this work we take the viewpoint that we cannot afford (or choose not to perform) the matrix-vector multiplication $y_{t+1} = Ax_t$ exactly. Among other applications, such a scenario naturally arises when the dimension of the matrix A is extremely large, so that even storage of the vector y_{t+1} (even in sparse format) is too expensive. For example, this is a common situation for FCI calculations in quantum chemistry since the dimension of the matrix A grows exponentially with respect to the number of electrons in the chemical system.

Thus, in power iteration, we would replace the matrix multiplication step by a map:

$$y_{t+1} = F_m(A, x_t). \tag{1}$$

Here, given the matrix A and the current iterate x_t, the map F_m, either deterministic or stochastic, outputs an approximation of the product $y_{t+1} \approx Ax_t$. Different choices of F_m correspond to various recently proposed algorithms, as will be discussed below. We have used the index m to indicate the "complexity" (computational cost) of F_m; the specific meaning depends on the choice of the family of maps. Replacing the matrix-vector multiplication by (1), we get the inexact power iteration algorithm (Algorithm 1a) and its unnormalized version (Algorithm 1b).

Notice that the two versions of inexact power iteration algorithm are equivalent if the function $F_m(A, \cdot)$ is homogeneous; we will make this a standing assumption in our analysis.

Algorithm 1a: Inexact Power Iteration

```
Initialization: Choose a normalized vector $x_0 \in \mathbb{R}^N$, $\|x_0\|_2 = 1$, $u_1^\top x_0 \neq 0$.
for $t = 0, 1, 2, \ldots$, while not converged do
    $y_{t+1} = F_m(A, x_t)$;
    $x_{t+1} = y_{t+1}/\|y_{t+1}\|_2$;
end
```

Algorithm 1b: Inexact Power Iteration without Normalization

```
Initialization: Choose a vector $v_0 \in \mathbb{R}^N$, $u_1^\top v_0 \neq 0$.
for $t = 0, 1, 2, \ldots$, while not converged do
    $v_{t+1} = F_m(A, v_t)$;
end
```

Various inexact matrix-vector multiplication methods have been proposed in the literature for configuration interaction calculations, either deterministic or stochastic; see, e.g., earlier attempts in [15, 6, 13, 16, 8, 11, 12, 17, 1], the FCIQMC approach [5, 4, 7, 3, 27], the semistochastic approach [23, 2, 14, 28], other stochastic approaches [30, 10, 19], and various deterministic strategies for compressed or truncated representation of the wave functions [24, 25, 21, 9, 18, 32, 31, 20, 26, 34, 33]. In this work, we will focus on two such strategies: full configuration interaction quantum Monte Carlo (FCIQMC) [5, 7] and fast randomized iteration (FRI) [19]. In some sense, these methods represent two ends of the spectrum of the possibilities, and so the analysis of those can be easily extended to other methodologies. FCIQMC uses interacting particles to represent the vector v_t and stochastic evolution of particles to represent the action of the matrix A on the vector v_t. FRI, on the other hand, is based on exact matrix-vector multiplication and stochastic schemes to compress the resulting vectors into sparse ones with a given number of nonzero entries. These algorithms will be discussed and analyzed in section 3, following the general analysis framework we establish in section 2.

The rest of the paper is organized as follows. In section 2, we provide a convergence analysis for a generic class of inexact power iteration. In section 3, we give more details of FCIQMC and FRI and analyze them following the convergence analysis established in section 2. In section 4, we perform numerical tests on the 2D Hubbard model and some chemical molecules to compare the various algorithms and to verify the analysis results.

2. General convergence analysis of the inexact power iteration.

An advantage of taking a unified framework of various algorithms is that their convergence can be understood in a fairly generic way, which also facilitates comparisons of different proposed strategies. In this section, we establish a general convergence theorem of the inexact power iteration.

The convergence of the iteration to the desired eigenvector will be measured by the angle of the vectors. Recall that the angle between two vectors v and w is given by

$$\theta(v, w) = \arccos\left(\frac{|\langle v, w\rangle|}{\|v\|_2\|w\|_2}\right). \tag{2}$$

From the definition, it is obvious that $\theta(v,w) = \theta(av,bw)$ for any vectors v,w and real numbers a,b. In view of this insensitivity of the constant multiple of vectors in the error measure, if the inexact matrix-vector multiplication $F_m(A,v_t)$ satisfies the homogeneity assumption below, the two versions of inexact power iterations with or without normalization (Algorithms 1a and 1b) are equivalent.

Assumption 1 (homogeneity).

$$F_m(A,cv) = cF_m(A,v) \tag{3}$$

for all vectors $v \in \mathbb{R}^N$ and real number $c \in \mathbb{R}$.

More precisely, if the initial vectors of the two algorithms are the same up to a number $x_0 = c_0 v_0$, then there exist numbers c_t such that $v_t = c_t x_t$ for all t. Therefore, $\theta(u_1,v_t) = \theta(u_1,x_t)$. In the following, when we analyze the algorithm, we will always use v_t for the unnormalized iterate and x_t for the normalized version $x_t = v_t/\|v_t\|_2$.

In analyzing the effect of the inexact matrix-vector multiplication, we write $F_m(A,v_{t-1})$ as a sum of the exact matrix-vector product with an error term

$$v_t = F_m(A,v_{t-1}) = Av_{t-1} + \xi_t, \tag{4}$$

where ξ_t is the error of the inexact multiplication at step t, and the dependence on m is suppressed to keep the notation simple. Note that ξ_t can be either deterministic or stochastic depending on the choice of F_m. For example, ξ_t is deterministic for the hard thresholding compression and stochastic for both the FCIQMC and FRI methods. While we will proceed viewing v_t as a stochastic process, the results apply to the deterministic case as well.

Denote by $\mathcal{F}_t = \sigma(v_1,v_2,\ldots,v_t)$ the σ-algebra generated by $v_1,v_2,\ldots,v_t$. We assume that error ξ_t satisfies the following properties. Note that this assumption holds for both the FCIQMC and FRI algorithms, as we will prove in section 3.

Assumption 2. *The error ξ_t in the inexact matrix-vector product* (4) *satisfies the following:*

(a) *Martingale difference sequence condition,*

$$\mathbb{E}\left(\xi_t \mid \mathcal{F}_{t-1}\right) = 0. \tag{5}$$

(b) *Expectation 2-norm bound,*

$$\mathbb{E}\left(\|\xi_t\|_2^2 \mid \mathcal{F}_{t-1}\right) \le C_e \frac{\|A\|_1^2 \|v_{t-1}\|_1^2}{m}, \tag{6}$$

where C_e is a constant that is scale invariant of A (i.e., it does not depend on the norm of A).

(c) *Growth of expectation 1-norm bound,*

$$\mathbb{E}\left(\|v_t\|_1 \mid \mathcal{F}_{t-1}\right) \le \|A\|_1 \|v_{t-1}\|_1. \tag{7}$$

To fully appreciate Assumption 2, a few remarks are in order. The martingale difference sequence property is just assumed here for convenience; in fact, the convergence result extends to the biased case, as we will see in Theorem 2. The other two assumptions are more essential. Assumption 2(b) indicates that the error of the inexact matrix-vector product $F_m(A,v_{t-1})$ can be controlled by the sparsity of v_{t-1}, as the 1-norm of v_{t-1} is a sparsity measure. This is a natural assumption, considering that the compression of a vector would be easier if the vector is more sparse. The bound depends proportionally on the inverse of m, so that one could control the error of the inexact matrix-vector multiplication at the price of increasing the complexity. Note that $1/m$ dependence can be understood as a standard Monte Carlo error scaling. More detailed discussions can be found in section 3 when specific algorithms are analyzed. Assumption 2(c) then assumes that the sparsity is not destroyed by the error in the iteration, since otherwise we would lose control of the accuracy of the inexact matrix-vector multiplication.

We now state the convergence theorem for the inexact power iteration algorithms, Algorithms 1a and 1b. The theorem provides a convergence guarantee with high probability given that the complexity parameter is sufficiently large, with the number of iteration steps T chosen properly. Note that the logarithmic dependence of T on the spectral gap λ_1/λ_2 and the error criteria δ and ε is expected from the standard power method. The dependence of m_0, the complexity parameter, on the ratio of the 1-norm and 2-norm of A is due to the competition between the 1- and 2-norm growth of the iterate, where the 1-norm matters for the control of the error of the inexact matrix-vector product.

THEOREM 1. *For inexact power iteration Algorithm* 1b*, under Assumption* 2*, for any precision* $\varepsilon > 0$ *and small probability* $\delta \in (0,1)$*, there exist time*

$$T = \log(\lambda_1/\lambda_2)^{-1} \log\left(\frac{2\sqrt{2}}{\sqrt{\delta}\varepsilon \cos\theta(u_1, v_0)}\right) \tag{8}$$

and measure of complexity

$$m_0 = \frac{4C_e}{\delta\varepsilon^2\big(\cos\theta(u_1,v_0)\big)^2}\frac{\|v_0\|_1^2}{\|v_0\|_2^2}T\bigg(\frac{\|A\|_1}{\|A\|_2}\bigg)^{2T}, \tag{9}$$

such that with probability at least $1-2\delta$*, for any* $m \geq m_0$*, it holds that*

$$\tan\theta(u_1, v_T) \leq \varepsilon. \tag{10}$$

Moreover, if Assumption 1 *is satisfied, the same result holds for Algorithm* 1a.

Before proving the theorem, let us collect a few immediate consequences of Assumption 2. The proof is obvious and will be omitted.

LEMMA 1. *If the error* ξ_t *satisfies Assumption* 2*, we have the following:*

(a) *The error is unbiased,*

$$\mathbb{E}\,\xi_t = 0. \tag{11}$$

(b) *The error at different step is uncorrelated, in particular,*

$$\mathbb{E}\,\xi_t^\top A^r \xi_s = 0 \tag{12}$$

for any $t \neq s$ *and for all nonnegative integers* r*.*

(c) *The expected* 2*-norm of the error can be controlled as*

$$\mathbb{E}\,\|\xi_t\|_2^2 \leq C_e\|A\|_1^{2t}\frac{\|v_0\|_1^2}{m}. \tag{13}$$

Proof of Theorem 1. From the iteration of Algorithm 1b, we obtain

$$\begin{aligned} v_t &= Av_{t-1} + \xi_t \\ &= A^t v_0 + A^{t-1}\xi_1 + \cdots + A\xi_{t-1} + \xi_t. \end{aligned}$$

Since the error ξ_t is unbiased, we have

$$\mathbb{E}\, v_t = A^t v_0.$$

We now control the variance of v_t. Since ξ_t is uncorrelated, we have

$$\mathbb{E}\, v_t^\top v_t = v_0^\top A^{2t} v_0 + \mathbb{E}\, \xi_1^\top A^{2t-2}\xi_1 + \cdots + \mathbb{E}\, \xi_{t-1}^\top A^2 \xi_{t-1} + \mathbb{E}\, \xi_t^\top \xi_t.$$

Thus,

$$\mathbb{E}\, v_t^\top v_t - \mathbb{E}\, v_t^\top \mathbb{E}\, v_t = \mathbb{E}\, \xi_1^\top A^{2t-2}\xi_1 + \cdots + \mathbb{E}\, \xi_{t-1}^\top A^2 \xi_{t-1} + \mathbb{E}\, \xi_t^\top \xi_t.$$

Using Lemma 1, we estimate

$$\begin{aligned} |\mathbb{E}\, v_t^\top v_t - \mathbb{E}\, v_t^\top \mathbb{E}\, v_t| &= |\mathbb{E}\, \xi_1^\top A^{2t-2}\xi_1 + \cdots + \mathbb{E}\, \xi_{t-1}^\top A^2 \xi_{t-1} + \mathbb{E}\, \xi_t^\top \xi_t| \\ &\le \mathbb{E}\, |\xi_1^\top A^{2t-2}\xi_1| + \cdots + \mathbb{E}\, |\xi_{t-1}^\top A^2 \xi_{t-1}| + \mathbb{E}\, |\xi_t^\top \xi_t| \\ &\le \lambda_1^{2t-2}\mathbb{E}\, |\xi_1^\top \xi_1| + \cdots + \lambda_1^2 \mathbb{E}\, |\xi_{t-1}^\top \xi_{t-1}| + \mathbb{E}\, |\xi_t^\top \xi_t| \\ &\le C_e \frac{\|A\|_1^2 \|v_0\|_1^2}{m} \Big(\lambda_1^{2t-2} + \cdots + \lambda_1^2 \|A\|_1^{2t-4} + \|A\|_1^{2t-2} \Big) \\ &= C_e \frac{\|A\|_1^2 \|v_0\|_1^2}{m} \frac{\|A\|_1^{2t} - \|A\|_2^{2t}}{\|A\|_1^2 - \|A\|_2^2}, \end{aligned}$$

where in the last step we used the fact that $\lambda_1 = \|A\|_2$ since λ_1 is the largest eigenvalue. Recall that for the symmetric matrix, $\|A\|_1 \ge \|A\|_2$, and hence we have

$$|\mathbb{E}\, v_t^\top v_t - \mathbb{E}\, v_t^\top \mathbb{E}\, v_t| \le C_e \frac{\|v_0\|_1^2}{m}\, t \|A\|_1^{2t}. \tag{14}$$

By analogous arguments, for

$$\mathbb{E}\, v_t v_t^\top - \mathbb{E}\, v_t \mathbb{E}\, v_t^\top = \mathbb{E}\, A^{t-1}\xi_1 \xi_1^\top A^{t-1} + \cdots + \mathbb{E}\, A\xi_{t-1}\xi_{t-1}^\top A + \mathbb{E}\, \xi_t \xi_t^\top,$$

we get

$$\big\| \mathbb{E}\, v_t v_t^\top - \mathbb{E}\, v_t \mathbb{E}\, v_t^\top \big\|_2 \le C_e \frac{\|v_0\|_1^2}{m}\, t \|A\|_1^{2t}. \tag{15}$$

Let us now estimate the angle between v_t and u_1—the eigenvector associated with the largest eigenvalue. By definition,

$$\big(\tan\theta(u_1, v_t)\big)^2 = \frac{\|v_t\|_2^2 - (u_1^\top v_t)^2}{(u_1^\top v_t)^2}. \tag{16}$$

For the denominator, we know the expectation,

$$\mathbb{E}\, u_1^\top v_t = u_1^\top A^t v_0 = \lambda_1^t (u_1^\top v_0),$$

and the variance,

$$\begin{aligned}\operatorname{Var}(u_1^\top v_t) &= u_1^\top(\mathbb{E}\, v_t v_t^\top - \mathbb{E}\, v_t \mathbb{E}\, v_t^\top)u_1\\ &\le \|\mathbb{E}\, v_t v_t^\top - \mathbb{E}\, v_t \mathbb{E}\, v_t^\top\|_2 \overset{(15)}{\le} C_e \frac{\|v_0\|_1^2}{m}\, t\|A\|_1^{2t}.\end{aligned}$$

The Chebyshev inequality implies that

$$\mathbb{P}\Big(\big|u_1^\top v_t - \lambda_1^t u_1^\top v_0\big| \ge \sqrt{\frac{C_e t}{m\delta}}\|v_0\|_1\|A\|_1^t\Big) \le \delta,$$

and hence, as $|\lambda_1^t u_1^\top v_0| - |u_1^\top v_t| \le |u_1^\top v_t - \lambda_1^t u_1^\top v_0|$,

$$\mathbb{P}\Big(\big|u_1^\top v_t\big| \le \big|\lambda_1^t u_1^\top v_0\big| - \sqrt{\frac{C_e t}{m\delta}}\|v_0\|_1\|A\|_1^t\Big) \le \delta,$$

or equivalently

$$\mathbb{P}\Big(\big|u_1^\top v_t\big|^2 \le \Big(\big|\lambda_1^t u_1^\top v_0\big| - \sqrt{\frac{C_e t}{m\delta}}\|v_0\|_1\|A\|_1^t\Big)^2\Big) \le \delta.$$

For the numerator of (16), the expectation is

$$\begin{aligned}\mathbb{E}\left(\|v_t\|_2^2 - (u_1^\top v_t)^2\right) &= \sum_{i=2}^N u_i^\top \mathbb{E}\, v_t v_t^\top u_i\\ &= \sum_{i=2}^N u_i^\top(\mathbb{E}\, v_t v_t^\top - \mathbb{E}\, v_t \mathbb{E}\, v_t^\top)u_i + \sum_{i=2}^N (u_i^\top \mathbb{E}\, v_t)^2\\ &\le \operatorname{tr}(\mathbb{E}\, v_t v_t^\top - \mathbb{E}\, v_t \mathbb{E}\, v_t^\top) + \lambda_2^{2t}\|v_0\|_2^2\\ &= (\mathbb{E}\, v_t^\top v_t - \mathbb{E}\, v_t^\top \mathbb{E}\, v_t) + \lambda_2^{2t}\|v_0\|_2^2\\ &\overset{(14)}{\le} C_e \frac{\|v_0\|_1^2}{m}\, t\|A\|_1^{2t} + \lambda_2^{2t}\|v_0\|_2^2.\end{aligned}$$

By the Markov inequality, for any $\delta \in (0,1)$

$$\mathbb{P}\left(\|v_t\|_2^2 - (u_1^\top v_t)^2 \ge \frac{1}{\delta}\Big(C_e\frac{\|v_0\|_1^2}{m}\, t\|A\|_1^{2t} + \lambda_2^{2t}\|v_0\|_2^2\Big)\right) \le \delta.$$

Therefore,

$$\mathbb{P}\left(\big(\tan\theta(u_1, v_t)\big)^2 \le \frac{1}{\delta}\frac{C_e\frac{\|v_0\|_1^2}{m}\, t\|A\|_1^{2t} + \lambda_2^{2t}\|v_0\|_2^2}{\Big(\big|\lambda_1^t u_1^\top v_0\big| - \sqrt{\frac{C_e t}{m\delta}}\|v_0\|_1\|A\|_1^t\Big)^2}\right) \ge 1 - 2\delta. \tag{17}$$

We can then explicitly check, where T and m_0 are defined as in (8) and (9), that for $m \ge m_0$, we have

$$\mathbb{P}\big(\tan\theta(u_1, v_T) \le \varepsilon\big) \ge 1 - 2\delta$$

and thus the claim of the theorem. ∎

As mentioned above, it is possible to drop the martingale difference sequence condition in Assumption 2 and get a similar result. The reason is that the second moment bound (6) can be used to control the bias of ξ_t. We state this as the following theorem.

THEOREM 2. *For inexact power iteration Algorithm* 1b*, under Assumptions* 2(b) *and* 2(c)*, for any precision $\varepsilon > 0$ and small probability $\delta \in (0,1)$, there exist time*

$$T = \log(\lambda_1/\lambda_2)^{-1} \log\left(\frac{4}{\sqrt{\delta}\varepsilon \cos\theta(u_1, v_0)}\right) \tag{18}$$

and measure of complexity

$$m_0 = \frac{8C_e}{\delta\varepsilon^2\big(\cos\theta(u_1,v_0)\big)^2}\frac{\|v_0\|_1^2}{\|v_0\|_2^2}\,T^2\left(\frac{\|A\|_1}{\|A\|_2}\right)^{2T}, \tag{19}$$

such that with probability at least $1-2\delta$, for any $m \ge m_0$, it holds that

$$\tan\theta(u_1, v_T) \le \varepsilon. \tag{20}$$

Moreover, if Assumption 1 *is satisfied, the same result holds for Algorithm* 1a.

Proof. Note that

$$\begin{aligned} u_1^\top v_t &= u_1^\top A^t v_0 + u_1^\top A^{t-1}\xi_1 + \cdots + u_1^\top A\xi_{t-1} + u_1^\top \xi_t \\ &= \lambda_1^t u_1^\top v_0 + \lambda_1^{t-1}u_1^\top \xi_1 + \cdots + \lambda_1 u_1^\top \xi_{t-1} + u_1^\top \xi_t, \end{aligned}$$

so we get

$$\begin{aligned} \mathbb{E}\,(u_1^\top v_t - \lambda_1^t u_1^\top v_0)^2 &= \mathbb{E}\,(\lambda_1^{t-1}u_1^\top\xi_1 + \cdots + \lambda_1 u_1^\top \xi_{t-1} + u_1^\top\xi_t)^2 \\ &= \sum_{i,j=1}^t \lambda_1^{2t-i-j}\mathbb{E}\,u_1^\top \xi_i\xi_j^\top u_1 \\ &\le \sum_{i,j=1}^t \lambda_1^{2t-i-j}\big(\mathbb{E}\,\|\xi_i\|_2^2\mathbb{E}\,\|\xi_j\|_2^2\big)^{1/2} \\ &\le C_e t^2\|A\|_1^{2t}\frac{\|v_0\|_1^2}{m}. \end{aligned}$$

Moreover,

$$\begin{aligned} \mathbb{E}\,\big(\|v_t\|_2^2 - (u_1^\top v_t)^2\big)^2 &= \sum_{i=2}^N (u_i^\top A^t v_0)^2 + 2\sum_{i=2}^N\sum_{b=1}^t \mathbb{E}\,u_i^\top A^t v_0 \xi_b^\top A^{t-b}u_i \\ &\quad + \sum_{i=2}^N\sum_{a=1}^t\sum_{b=1}^t \mathbb{E}\,u_i^\top A^{t-a}\xi_a\xi_b^\top A^{t-b}u_i \\ &\le \lambda_2^{2t}\|v_0\|_2^2 + 2t\sqrt{C_e}\lambda_2^t\|v_0\|_2\|A\|_1^t\frac{\|v_0\|_1}{\sqrt{m}} + t^2C_e\|A\|_1^{2t}\frac{\|v_0\|_1^2}{m} \\ &\le 2\lambda_2^{2t}\|v_0\|_2^2 + 2t^2C_e\|A\|_1^{2t}\frac{\|v_0\|_1^2}{m}, \end{aligned}$$

where the Cauchy–Schwarz inequality is used in the last line. Thus we can again use the Markov inequality to bound both numerator and denominator on the right-hand side of (16) to obtain the claimed result. ∎

3. Algorithms.

In this section, we will review two stochastic power iteration methods recently proposed in the literature: full configuration interaction quantum Monte Carlo (FCIQMC) [5] and fast randomized iteration (FRI) [19]. They can be analyzed in the same framework we established in the previous section. In particular, we prove the convergence of the two algorithms using Theorem 1. We focus on these two methods since in some sense they represent opposite ends of inexact matrix-vector multiplication strategies. It is possible to combine the ideas and get a menagerie of different approaches, which possibly yield better results; our analysis can be extended to these as well. We will also comment on two variants: iFCIQMC and hard thresholding (HT), closely related to the FCIQMC and FRI approaches.

Without loss of generality, we will assume matrix A is close to the identity matrix, and thus the eigenvalues λ_i are close to 1 (we can always scale and center the original matrix so that this is true).

3.1. Full configuration interaction quantum Monte Carlo.

3.1.1. Algorithm description. FCIQMC is an algorithm originating from the quantum chemistry literature and used to calculate the ground state energy of a many-body electron system by a Monte Carlo algorithm for the full configuration interaction of the many-body Hamiltonian [5].

Let the Hamiltonian be a real symmetric matrix $H \in \mathbb{R}^{N\times N}$ under the Slater determinant basis. To find the ground state (the smallest eigenvector) of H, we write $A = I - \delta H$ for $\delta > 0$ sufficiently small and hence focus on the largest eigenvalue of A; this can be viewed as a first order truncation of the Taylor series of $e^{-\delta H}$. It is also possible to construct other variants of A from H, which we will not do here.

FCIQMC can be viewed as a stochastic inexact power iteration for finding the largest eigenvector of A, which corresponds more naturally to the unnormalized version of the inexact power iteration (Algorithm 1b). In the algorithm, the vector v_t is not stored as a vector, but represented as a collection of "signed particles" $\{\alpha_t^{(i)}\}_{i=1}^{M_t}$, where M_t is the number of signed particles at iteration step t. Each signed particle α has two attributes: location $l_\alpha \in \{1,2,\dots,N\}$ and sign $s_\alpha \in \{1,-1\}$. Denote by $e_l \in \mathbb{R}^N$ the standard basis vector with value 1 at its lth component and 0 at every other component. Then each signed particle α represents a signed unit vector $\alpha = s_\alpha e_{l_\alpha}$. The vector $v_t \in \mathbb{Z}^N$ is given by the sum of all signed particles at time t:

$$v_t = \sum_{i=1}^{M_t} \alpha_t^{(i)}. \tag{21}$$

With some ambiguity of notation, we refer to both the set of particles and the corresponding vector as v_t, connected by (21). As we always assume that the particles with opposite signs on the same location are annihilated (see the annihilation step in the algorithm description below), the vector v_t uniquely determines the set of particles.

In FCIQMC, the inexact matrix-vector multiplication $F_m(A, v_t)$ consists of three steps of particle evolution: spawning, diagonal death/cloning, and annihilation. Write $A = A_d + A_o$, with A_d the diagonal part and A_o the off-diagonal part. The spawning step approximates $A_o v_t$; the diagonal death/cloning step approximates $A_d v_t$, and the annihilation step sums up the results from the previous two steps and approximates the summation $Av_t = A_o v_t + A_d v_t$. The three steps will be described in more detail below.

Spawning. Each signed particle α (we suppress the index of $\alpha_t^{(i)}$ to simplify notation) is allowed to spawn a child particle to another location, corresponding to

```
Algorithm 2a: FCIQMC - Spawning
  Input  : Set of particles: $\{\alpha^{(i)}\}_{i=1,\ldots,M}$, matrix $A$
  Output: New set of particles after spawning: $\{\alpha^{(j),\,\mathrm{sp}}\}_{j=1,\ldots,M^{\mathrm{sp}}}$
  $M^{\mathrm{sp}} = 0$;
  for each particle $\alpha \in \{\alpha^{(i)}\}_{i=1,\ldots,M}$ do
      Select a spawning location $l$ with probability $p_{\mathrm{loc}}(l \mid l_\alpha)$;
      Determine the expected number of children
          $Q = \frac{|A_o(l, l_\alpha)|}{p_{\mathrm{loc}}(l \mid l_\alpha)};$
      Randomly choose the number of children
          $n = \begin{cases} \lfloor Q \rfloor & \text{w.p.} \quad 1-(Q-\lfloor Q \rfloor), \\ \lfloor Q \rfloor + 1 & \text{w.p.} \quad Q - \lfloor Q \rfloor; \end{cases}$
      Assign the sign of each children as $s = \mathrm{sgn}(A_o(l, l_\alpha)s_\alpha) = \mathrm{sgn}\big((A_o\alpha)(l)\big)$;
      Increase the number of particles $M^{\mathrm{sp}} = M^{\mathrm{sp}} + n$;
      Add $n$ particles with location $l$ and sign $s$ into the spawning set $\{\alpha^{(j),\mathrm{sp}}\}$.
  end
```

a nonzero component of $A_o\alpha = s_\alpha A_o(:, l_\alpha)$.[1] The location of spawning is chosen at random, with probability $p_{\mathrm{loc}}(l \mid l_\alpha)$, which is chosen in the original FCIQMC algorithm to be uniformly random over all nonzero components of $A_o\alpha$ for some simple Hamiltonian H. In general, $p_{\mathrm{loc}}(\cdot \mid l_\alpha)$ can be more complicated; we refer readers to [5] for more details. In the remainder of the paper, $p_{\mathrm{loc}}(\cdot \mid l_\alpha)$ is assumed to be a uniform distribution over all nonzero components of $A_o\alpha$, while our analysis can be extended to other choices of $p_{\mathrm{loc}}(\cdot \mid l_\alpha)$.

Once the location l is chosen, n (possibly 0) children particles are spawned with the same sign $s = \mathrm{sgn}(A_o(l, l_\alpha)s_\alpha)$ determined by the sign of vector entry $(A_o\alpha)(l)$ and the particle α. The location l and number n are stochastically chosen such that the overall step gives an unbiased estimate of $A_o\alpha$:

$$\mathbb{E}\left(nse_l \mid \alpha\right) = A_o\alpha. \tag{22}$$

Please refer to Algorithm 2a for details.

Diagonal cloning/death. This step represents $A_d v_t$ as a collection of particles in an analogous way to the spawning step. For every signed particle α, we would consider children particles on the location l_α (i.e., the location of the new particles is chosen to be l_α) and obtain an unbiased representation

$$\mathbb{E}\left(nse_{l_\alpha} \mid \alpha\right) = A_d\alpha. \tag{23}$$

The details can be found in Algorithm 2b; the key steps are similar to Algorithm 2a.

Annihilation. The annihilation step merges the children particles from the previous two steps and removes all pairs of particles with the same location and opposite

[1] MATLAB notation $A(:, l)$ is used to denote the lth column of A.

Algorithm 2b: FCIQMC - Diagonal cloning/death

```
Input  : Set of particles: $\{\alpha^{(i)}\}_{i=1,\dots,M}$, matrix $A$
Output: New set of particles after diagonal cloning / death:
        $\{\alpha^{(j),\,\mathrm{diag}}\}_{j=1,\dots,M^{\mathrm{diag}}}$
$M^{\mathrm{diag}} = 0$;
for each particle $\alpha \in \{\alpha^{(i)}\}_{i=1,\dots,M}$ do
    Determine the expected number of children at $l_\alpha$
        $Q = |A(l_\alpha, l_\alpha)|$;
    Randomly choose the number of children
        $n = \begin{cases} \lfloor Q \rfloor & \text{w.p.} \quad 1 - (Q - \lfloor Q \rfloor), \\ \lfloor Q \rfloor + 1 & \text{w.p.} \quad Q - \lfloor Q \rfloor; \end{cases}$
    Assign the sign of each children as $s = \mathrm{sgn}\big((A_d\alpha)(l_\alpha)\big)$;
    Increase the number of particles $M^{\mathrm{diag}} = M^{\mathrm{diag}} + n$;
    Add $n$ particles with location $l_\alpha$ and sign $s$ into the set $\{\alpha^{(j),\mathrm{diag}}\}$.
```

signs. If we denote by v^{sp} and v^{diag} the corresponding vector representation of the particles in the spawning and diagonal cloning/death steps, the annihilation steps create a collection of particles representing the new vector $v = v^{\mathrm{sp}} + v^{\mathrm{diag}}$. Applying the three steps above to the particles representing v_t, we obtain the new set of particles v_{t+1} at time $t+1$. Since by construction

$$\mathbb{E}\left(v^{\mathrm{sp}} \mid v_t\right) = A_o v_t \quad \text{and} \quad \mathbb{E}\left(v^{\mathrm{diag}} \mid v_t\right) = A_d v_t,$$

we have on expectation

$$\mathbb{E}\left(v_{t+1} \mid v_t\right) = A v_t. \tag{24}$$

In terms of the notation used in the framework of the inexact power iteration, v_{t+1} represented using particles can be viewed as the approximate matrix-vector product $F_m(A, v_t)$:

$$F_m(A, v_t) := v_{t+1} = \sum_{i=1}^{M_{t+1}} \alpha_{t+1}^{(i)} = A v_t + \xi_{t+1}, \tag{25}$$

where ξ_{t+1} is introduced in the last equality to denote the error from the approximate matrix-vector multiplication through the stochastic particle representation. As we will show in the analysis below, the accuracy of the FCIQMC iteration is controlled by the number of particles M_t, and thus it plays the role of the complexity parameter m in our general framework. We will drop the subscript m for F_m in what follows for FCIQMC, as the complexity parameter is implicit.

Now that we have defined the inexact matrix-vector multiplication $F(A, v_t)$ in FCIQMC, we may apply this in the inexact power iteration as Algorithm 1b. However, this can be problematic in practice. Recall that $A = I - \delta H$ is assumed to be a perturbation of identity, so its eigenvalue is around 1. If the largest eigenvalue of A is strictly larger than 1, when the signed particles become a good approximation to the leading eigenvector, the number of particles M_t will grow exponentially with rate λ_1, which quickly increases the computational cost and memory requirement. It is also possible (while the probability is tiny) that the number of particles may decrease to 0 due to randomness.

In practice, it is desirable to maintain controls of the number of particles to make the algorithm more stable. One such strategy is to introduce a shift $s_t \in \mathbb{R}$ and use matrix

$$\widetilde{A} = A + \delta s_t I = I - \delta(H - s_t I) \tag{26}$$

instead of A at the tth step. Notice that s_t only shifts the eigenvalues while not changing the eigenspace. The shift s_t is adjusted dynamically to control the number of particles. With such shifts, the full FCIQMC algorithm is presented in Algorithm 2.

Algorithm 2: FCIQMC

```
Initialization: $t = 0$ and set initial particles $v_0$.
while $M_t \le M^{target}$, the target population do
    // Phase 1: FCIQMC with fixed shift $s_0$
    Spawning step: Use Algorithm 2a with $v_t$ and $A + \delta s_0 I$ to get particle set
      $v^{\mathrm{sp}}$;
    Diagonal death/cloning step: Use Algorithm 2b with $v_t$ and $A + \delta s_0 I$ to
      get particle set $v^{\mathrm{diag}}$;
    Annihilation step to get the particle set of the next time step
      $v_{t+1} = v^{\mathrm{sp}} + v^{\mathrm{diag}}$;
    Update $M_t$ and set $t = t + 1$;
end
Set $s_t = s_0$;   // Initialize the dynamic shift
while $t < t_{\max}$ do
    // Phase 2: FCIQMC with dynamic shift $s_t$
    Spawning step: Use Algorithm 2a with $v_t$ and $A + \delta s_t I$ to get particle set
      $v^{\mathrm{sp}}$;
    Diagonal death/cloning step: Use Algorithm 2b with $v_t$ and $A + \delta s_t I$ to
      get particle set $v^{\mathrm{diag}}$;
    Annihilation step: Merge the two sets of particles $v_{t+1} = v^{\mathrm{sp}} + v^{\mathrm{diag}}$;
    Update the shift $s_t$ as

      (27)    $s_t = \begin{cases} s_{t-q} - \frac{\eta}{q}\big(\ln M_t - \ln M_{t-q}\big) & \text{if } t = 0 \pmod{q}, \\ s_{t-1} & \text{otherwise;} \end{cases}$

    Update $M_t$ and set $t = t + 1$;
end
```

Algorithm 2 contains two phases for different strategies of choosing the shifts and thus controlling the particle population. In Phase 1, the shift is fixed to be s_0, which is chosen such that $|A(i,i) - s_0| \ge 1$ for all i so that the particle number is most likely to grow exponentially till the target population M^{target}. In the second phase, the shift is dynamically adjusted, so as to control the growth of the population by a negative feedback loop. The target number of population M^{target} is chosen to be sufficiently large that the variance is small enough to ensure convergence. It plays the role as the "complexity" m in Theorem 1. η and q are two parameters to control the fluctuation of the number of particles. For details on the parameter choices, we refer the reader to the original paper on FCIQMC [5].

Energy estimator. Several estimators can be used to estimate the smallest eigenvalue of H based on the FCIQMC Algorithm 2, which is just a linear transformation of the largest eigenvalue λ_1 of A. One estimator is simply the shift s_t. When the algorithm converges, v_t is approximately proportional to the eigenvector u_1. Since s_t is adjusted to make the number of particles steady, the largest eigenvalue of $A+\delta s_t I$ is approximately 1, hence connecting s_t with the desired eigenvalue estimate; cf. (26). The other estimator we will consider is the projected energy estimator:

$$E_t = \frac{v_*^\top H v_t}{v_*^\top v_t}.$$

Here v_* is some fixed vector—for example, the Hartree–Fock state of the system. It is clear that when v_t becomes a good approximation of the eigenvector u_1, E_t gives a good estimate of the leading eigenvalue. In the numerical examples, we will focus on the projected energy estimator, since it can be applied to all algorithms we consider in this work (while the shift estimator is unique to FCIQMC, in practice, it gives results similar to those of the projected energy estimator).

3.1.2. Convergence analysis. Since FCIQMC can be viewed as an inexact power iteration as in (25), we apply Theorem 1 to analyze the convergence of FCIQMC. For simplicity, we will focus on the case that the shift is constantly 0, $s_t = 0$, since the shift does not affect the eigenvector, which is the main focus of Theorem 1. The probability distribution in the spawning step $p_{\text{loc}}(\cdot \mid l_\alpha)$ is assumed to be uniform distribution over all the nonzero entries of $A_o(:, l_\alpha)$. To avoid some degenerate cases, we will assume that each diagonal entry of A is nonzero and each column of A has more than 2 nonzero entries (so there is at least one possible location for children particles in the spawning step).

We now check the three conditions in Assumption 2. The unbiasedness is guaranteed by construction as discussed above for the FCIQMC algorithm; we have

$$\mathbb{E}\left(v_{t+1} \mid \mathcal{F}_t\right) = A v_t, \tag{28}$$

or equivalently, the error ξ_t is a martingale difference sequence:

$$\mathbb{E}\left(\xi_{t+1} \mid \mathcal{F}_t\right) = 0. \tag{29}$$

The expectation 2-norm bound is established in the following proposition.

Proposition 2. *For the inexact matrix-vector multiplication* (25) *in FCIQMC Algorithm* 2*, the error ξ_t satisfies*

$$\mathbb{E}\left(\|\xi_{t+1}\|_2^2 \mid \mathcal{F}_t\right) \le \left(\max_{1\le k\le n} (\|a_k\|_0 - 2)\|a_{o,k}\|_2^2 + \frac{1}{2}\right) \frac{\|v_t\|_1^2}{M_t}, \tag{30}$$

where $a_k = A(:, k)$ is the kth column vector of A, and $a_{o,k}$ is the kth column vector of A_o; thus $a_{o,k}$ equals a_k except for the kth entry $a_{o,k}(k) = 0$.

Proof. Since each particle evolves independently,

$$F(A, v_t) = F\left(A, \sum_{i=1}^{M_t} \alpha_t^{(i)}\right) = \sum_{i=1}^{M_t} F(A, \alpha_t^{(i)}).$$

Moreover $F(A, \alpha_t^{(i)})$ and $F(A, \alpha_t^{(j)})$ are independent for $i \neq j$ conditioned on $\mathcal{F}_t$.

By construction, $F(A, \alpha_t^{(i)})$ is unbiased, i.e.,

$$\mathbb{E}\left(F(A, \alpha_t^{(i)}) - A\alpha_t^{(i)} \mid \mathcal{F}_t\right) = 0.$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left(\|\xi_{t+1}\|_2^2 \mid \mathcal{F}_t\right) &= \mathbb{E}\left(\left\|\sum_{i=1}^{M_t}(F(A, \alpha_t^{(i)}) - A\alpha_t^{(i)})\right\|_2^2 \mid \mathcal{F}_t\right) \\
&= \mathbb{E}\left(\left(\sum_{i=1}^{M_t} F(A, \alpha_t^{(i)}) - A\alpha_t^{(i)}\right)^\top \left(\sum_{j=1}^{M_t} F(A, \alpha_t^{(j)}) - A\alpha_t^{(j)}\right) \mid \mathcal{F}_t\right) \\
&= \sum_{i=1}^{M_t} \mathbb{E}\left((F(A, \alpha_t^{(i)}) - A\alpha_t^{(i)})^\top (F(A, \alpha_t^{(i)}) - A\alpha_t^{(i)}) \mid \mathcal{F}_t\right) \\
&\quad + 2 \sum_{1 \le i < j \le M_t} \mathbb{E}\left((F(A, \alpha_t^{(i)}) - A\alpha_t^{(i)})^\top \mid \mathcal{F}_t\right) \mathbb{E}\left((F(A, \alpha_t^{(j)}) - A\alpha_t^{(j)}) \mid \mathcal{F}_t\right) \\
&= \sum_{i=1}^{M_t} \mathbb{E}\left(\left\|F(A, \alpha_t^{(i)}) - A\alpha_t^{(i)}\right\|_2^2 \mid \mathcal{F}_t\right).
\end{aligned}$$

Hence, it suffices to consider each particle individually. To simplify the notation, without loss of generality, let us consider a particle with $\alpha_t^{(i)} = e_k$ for some k. Since the spawning and diagonal cloning/death steps are independent and unbiased, we have the decomposition

$$\mathbb{E}\left(\|F(A, e_k) - Ae_k\|_2^2 \mid \mathcal{F}_t\right) = \mathbb{E}\left(\|F(A_o, e_k) - A_o e_k\|_2^2 \mid \mathcal{F}_t\right) + \mathbb{E}\left(\|F(A_d, e_k) - A_d e_k\|_2^2 \mid \mathcal{F}_t\right).$$

For the spawning step, since $A_o e_k = a_{o,k}$, there are $\|a_{o,k}\|_0$ locations to spawn. Recall that $p_{\mathrm{loc}}(\cdot \mid k)$ is assumed to be a uniform distribution, so each location is chosen with probability $\frac{1}{\|a_{o,k}\|_0}$. Following Algorithm 2a, we calculate that

$$F(A_o, e_k) = \begin{cases} \lfloor \|a_{o,k}\|_0 \, |a_k(j)| \rfloor \operatorname{sgn}(a_k(j)) e_j \\ \qquad \text{w.p.} \quad \left(1 - \left(\|a_{o,k}\|_0 \, |a_k(j)| - \lfloor \|a_{o,k}\|_0 \, |a_k(j)| \rfloor\right)\right) / \|a_{o,k}\|_0 \,, \\ \left(\lfloor \|a_{o,k}\|_0 \, |a_k(j)| \rfloor + 1\right) \operatorname{sgn}(a_k(j)) e_j \\ \qquad \text{w.p.} \quad \left(\|a_{o,k}\|_0 \, |a_k(j)| - \lfloor \|a_{o,k}\|_0 \, |a_k(j)| \rfloor\right) / \|a_{o,k}\|_0 \end{cases}$$

for each j such that $a_{o,k}(j) \neq 0$. Straightforward calculation yields

$$\begin{aligned}
\mathbb{E}\, \|F(A_o, e_k) - A_o e_k\|_2^2 &= \left(\|a_{o,k}\|_0 - 1\right) \|a_{o,k}\|_2^2 \\
&\quad + \frac{1}{\|a_{o,k}\|_0} \sum_{j,\, a_{o,k}(j) \neq 0} \left(\|a_{o,k}\|_0 \, a_k(j) - \lfloor \|a_{o,k}\|_0 \, a_k(j) \rfloor\right) \\
&\qquad \times \left(1 - \left(\|a_{o,k}\|_0 \, a_k(j) - \lfloor \|a_{o,k}\|_0 \, a_k(j) \rfloor\right)\right) \\
&\le \left(\|a_{o,k}\|_0 - 1\right) \|a_{o,k}\|_2^2 + \frac{1}{4}.
\end{aligned}$$

For the diagonal cloning/death step, we have

$$F(A_d, e_k) = \begin{cases} \lfloor |a_k(k)| \rfloor \operatorname{sgn}(a_k(k)) e_k & \text{w.p.} \quad 1 - \left(|a_k(i)| - \lfloor |a_k(i)| \rfloor\right), \\ (\lfloor |a_k(k)| \rfloor + 1) \operatorname{sgn}(a_k(k)) e_k & \text{w.p.} \quad |a_k(i)| - \lfloor |a_k(i)| \rfloor. \end{cases}$$

Therefore

$$\mathbb{E}\left(\|F(A_d, e_k) - A_d e_k\|_2^2 \mid \mathcal{F}_t\right) = \left(|a_k(i)| - \lfloor |a_k(i)| \rfloor\right)\Big(1 - \left(|a_k(i)| - \lfloor |a_k(i)| \rfloor\right)\Big) \leq \frac{1}{4}.$$

Summing up the contribution from the two steps, we arrive at

$$\mathbb{E}\left(\|F(A, e_k) - A e_k\|_2^2 \mid \mathcal{F}_t\right) \leq \left(\|a_k\|_0 - 2\right) \|a_{o,k}\|_2^2 + \frac{1}{2},$$

where we used $\|a_{o,k}\|_0 = \|a_k\|_0 - 1$. Thus

$$\mathbb{E}\left(\|F(A, v_t) - A v_t\|_2^2 \mid \mathcal{F}_t\right) \leq M_t \Big(\max_{1 \leq k \leq n} \left(\|a_k\|_0 - 2\right) \|a_{o,k}\|_2^2 + \frac{1}{2} \Big).$$

Since $M_t = \|v_t\|_1$, we can rewrite the above estimate as

$$\mathbb{E}\left(\|F(A, v_t) - A v_t\|_2^2 \mid \mathcal{F}_t\right) \leq \frac{\|v_t\|_1^2}{M_t} \Big(\max_{1 \leq k \leq n} (\|a_k\|_0 - 2) \|a_{o,k}\|_2^2 + \frac{1}{2} \Big). \qquad \square$$

Here we emphasize the important role of the annihilation step in FCIQMC reflected in the error analysis above. Only with the annihilation step is $M_t = \|v_t\|_1$ true, so that the growth of error is controlled as in the last step of the proof. In general, without annihilation, the error will be exponentially larger, as $\frac{M_t}{\|v_t\|_1}$ grows exponentially even when v_t is close to the eigenvector u_1. Suppose v_t is approximately u_1. Then $v_{t+1} \approx \lambda_1 v_t$. Therefore, $\|v_{t+1}\|_1 \approx \|A\|_2 \|v_t\|_1$. However, for the number of particles M_t without annihilation, $M_{t+1} \approx \||A|\|_2 M_t$, where $|A|$ is the entrywise absolute value of A. To see this, let us denote by v_t^+ the vector represented by all the particles with positive sign and by $-v_t^-$ the vector represented by all the particles with negative sign. Then $v_t = v_t^+ - v_t^-$. Denote $\tilde{v}_t = v_t^+ + v_t^-$. Then $M_t = \|\tilde{v}_t\|_1$ without annihilation. We can easily check that $\tilde{v}_t$ evolves according to $\tilde{v}_{t+1} = |A| \tilde{v}_t$. So finally $\tilde{v}_t$ will converge to the eigenvector of $|A|$, and $M_{t+1} \approx \||A|\|_2 M_t$. Noticing that $\|A\|_2 \leq \||A|\|_2 \leq \|A\|_1$, we know $\frac{M_t}{\|v_t\|_1}$ grows exponentially at rate $\frac{\||A|\|_2}{\|A\|_2}$ after convergence. Therefore if the number of particles M_t has an upper bound, which is always true in practice due to computational resource constraint, $\|v_t\|_1$ will decay to zero exponentially, which means the algorithm will not converge to the correct eigenvector. Also note that if the spawning distribution $p_{\mathrm{loc}}(\cdot \mid l_\alpha)$ is not exactly a uniform distribution, then $\mathbb{E}\left(\|F(A_o, e_k) - A_o e_k\|_2^2 \mid \mathcal{F}_t\right)$ will be bound by another constant depending on A_o. Therefore the bound of $\mathbb{E}\left(\|\xi_{t+1}\|_2^2 \mid \mathcal{F}_t\right)$ in the proposition will only differ by a constant multiplier.

Compared with Assumption 2, we observe that the particle number M_t plays the role of the "complexity" parameter. The more particles we have, the smaller the error is. We have the following corollary, assuming the particle number is bounded from below by m

Corollary 3. *If the particle number satisfies* $M_t \geq m$,

$$\mathbb{E}\left(\|F(A, v_t) - A v_t\|_2^2 \mid \mathcal{F}_t, M_t \geq m\right) \leq C_e \frac{\|A\|_1^2 \|v_t\|_1^2}{m}, \tag{31}$$

where $C_e = \frac{\max_k(\|a_k\|_0-2)\|a_{o,k}\|_2^2+\frac{1}{2}}{\|A\|_1^2}$ *is a parameter scale-invariant of* A.

In summary, FCIQMC satisfies Assumption 2(b) as long as the particle number is not too small. Note that in practice the particle number can be controlled by the dynamic shift s_t to ensure that it does not drop below the required lower bound.

Assumption 2(c), the growth of expectation 1-norm bound, can also be checked easily, since we have

$$\begin{aligned}\mathbb{E}\left(\|v_{t+1}\|_1 \mid \mathcal{F}_t\right) &= \mathbb{E}\left(\left\|F\left(A, \sum_{i=1}^{M_t}\alpha_t^{(i)}\right)\right\|_1 \mid \mathcal{F}_t\right) = \mathbb{E}\left(\left\|\sum_{i=1}^{M_t}F(A,\alpha_t^{(i)})\right\|_1 \mid \mathcal{F}_t\right)\\ &\leq \sum_{i=1}^{M_t}\mathbb{E}\left(\|F(A,\alpha_t^{(i)})\|_1 \mid \mathcal{F}_t\right) = \sum_{i=1}^{M_t}\|A\alpha_t^{(i)}\|_1\\ &\leq \sum_{i=1}^{M_t}\|A\|_1\|\alpha_t^{(i)}\|_1 = \|A\|_1\|v_t\|_1.\end{aligned}$$

In conclusion, we have verified the assumptions of Theorem 1, and thus it can be applied for the convergence and error analysis of FCIQMC.

3.1.3. Remarks on *i*FCIQMC.

*i*FCIQMC (initiator FCIQMC) [7] is a modified version of FCIQMC. It can be viewed as a bias-variance trade-off strategy to reduce the computational cost and error of the FCIQMC approach by restricting the spawning step.

The n locations are divided into two sets: the initiators L_i and noninitiators L_n with $L_i \cap L_n = \emptyset$, $L_i \cup L_n = \{1, 2, \dots, N\}$. The rule of iFCIQMC is that any particle α at a noninitiator location $l_\alpha \in L_n$ is only allowed to spawn children particles at locations already occupied by some other particles. If α spawns particles to a location unoccupied, then the children particles are discarded. An exception rule is that if at least two particles at noninitiator locations spawn children particles with the same sign at one unoccupied location, then the children particles are kept. There are no restrictions for spawning steps for particles in initiators. In the case that all the locations are initiators, $L_n = \emptyset$, iFCIQMC reduces to FCIQMC.

The initiators L_i are chosen at the beginning according to some prior knowledge. The initiators are then updated at each step of iteration. Suppose $n_{i,thre} \in \mathbb{N}$ is a fixed threshold. As soon as the number of particles at a noninitiator location is greater than the threshold $n_{i,thre}$, then the location becomes an initiator. Intuitively, initiators are more important locations for the eigenvector since they are occupied by many particles. The restrictions on the spawning ability of noninitiators reduce the computational cost and the variance of the inexact matrix-vector product while introducing only a small bias since there are few particles on noninitiators. Therefore, iFCIQMC can be viewed as a variance control technique for FCIQMC.

3.2. Fast randomized iteration.

In this section, we provide a numerical analysis based on our general framework for the convergence of fast randomized iteration (FRI), recently proposed in the applied mathematics literature [19], inspired by FCIQMC-type algorithms. The basic idea of the FRI method is to first apply matrix A on the vector of the current iterate, and then employ a stochastic compression algorithm to reduce the resulting vector to a sparse representation. The original convergence analysis [19] uses a norm motivated by viewing the vectors as random measures. In comparison, as we have seen in the proof of Theorem 1, our viewpoint and analysis are closer in spirit to numerical linear algebra, in particular the standard convergence analysis of the power method.

3.2.1. Algorithm description. The fast randomized iteration (FRI) algorithm is based on a choice of the random compression function $\Phi_m : \mathbb{R}^N \to \mathbb{R}^N$, which maps a full vector v to a sparse vector $\Phi_m(v)$ with approximately only m nonzero components. The sparsity of $\Phi_m(v)$ reduces the storage cost of the vector and associated computational cost. To combine FRI with the inexact power iteration, define

$$F_m(A, v_t) = \Phi_m(Av_t) \tag{32}$$

in Algorithms 1a and 1b. The error is $\xi_{t+1} = \Phi_m(Av_t) - Av_t$.

Thus the FRI algorithm is completely characterized by the choice of compression function Φ_m, which we assume has the following properties. These are adaptations of Assumptions 1 and 2 in the context of a compression function.

Assumption 3. *For any vector $v \in \mathbb{R}^N$, the compression function Φ_m satisfies the following:*

(a) *Homogeneity: For all $c \in \mathbb{R}$,*

$$\Phi_m(cv) = c\Phi_m(v). \tag{33}$$

(b) *Unbiasedness:*

$$\mathbb{E}\left(\Phi_m(v) \mid v\right) = v. \tag{34}$$

(c) *Variance bound: For some constant C_Φ independent of m and v,*

$$\mathbb{E}\left(\|\Phi_m(v) - v\|_2^2 \mid v\right) \le C_\Phi \frac{\|v\|_1^2}{m}. \tag{35}$$

(d) *Expectation 1-norm bound:*

$$\mathbb{E}\left(\|\Phi_m(v)\|_1 \mid v\right) = \|v\|_1. \tag{36}$$

The compression function Φ_m introduced in [19] is as follows: For a given vector $v \in \mathbb{R}^N$, first we sort the entries as $|v(q_1)| \ge |v(q_2)| \ge \cdots \ge |v(q_N)|$, where $q : [N] \to [N]$ is a permutation. The compression function consists of two parts. In the first part, large components of the vector are preserved exactly. Define

$$\tau = \max_{1 \le i \le N} \left\{ i : |v(q_i)| \ge \frac{\sum_{j=i}^{N} |v(q_j)|}{m + 1 - i} \right\},$$

with the convention $\max\{\emptyset\} = 0$, so $0 \le \tau \le m$. The compression function keeps the entries $v(q_i)$ for any $1 \le i \le \tau$,

$$\big(\Phi_m(v)\big)(q_i) = v(q_i) \qquad \forall i \le \tau.$$

Note that if $\|v\|_0 \le m$, all components are "large" and $\Phi_m(v) = v$; the input vector is kept without compression. The remaining $n - \tau$ components are considered "small." Under the compression we only keep a few entries, so the resulting vector $\Phi_m(v)$ has about m nonzero entries, as in Algorithm 3; the details are further discussed below.

In the second part of Algorithm 3, the set $B = \{q_{\tau+1}, q_{\tau+2}, \ldots, q_N\}$ consists of the indices of all "small" components to be compressed. Note that for the integer

```
Algorithm 3: FRI - compression function $\Phi_m$
    Input  : $v \in \mathbb{R}^N$, sparsity parameter $m$
    Output: $V = \Phi_m(v) \in \mathbb{R}^N$
    // Part 1:  Keep large components
    $B = \{1, 2, \dots, N\}$;
    $s = \|v\|_1$;
    $i' = \arg\max_{i \in B} |v(i)|$;
    $\tau = 0$;
    while $|v(i')| \geq \frac{s}{m - \tau}$ do
        $V(i') = v(i')$;
        $s = s - |v(i')|$;
        $\tau = \tau + 1$;
        $B = B \backslash \{i'\}$;
        $i' = \arg\max_{i \in B} |v(i)|$;
    end
    // Part 2:  Compress small components
    for each $i \in B$ do
        Generate nonnegative random integer $\{N_i\}$, such that
            $\mathbb{E}\, N_i = \frac{m - \tau}{s} |v(i)|$;
        $V(i) = \operatorname{sgn}(v(i)) N_i \frac{s}{m - \tau}$;
    end
```

random variable N_i, $i \in B$, only its expectation $\mathbb{E}\, N_i \in (0, 1)$ is specified, so there is still freedom to choose the probability distribution of $\{N_i\}_{i \in B}$. Here we only discuss independent Bernoulli (which is easy to understand) and systematic sampling (which we use in the numerical examples) approaches, while other choices are possible. Let us focus on the entries in B and define $v' \in \mathbb{R}^n$ such that $v'(i) = v(i)\mathbf{1}_{\{i \in B\}}$. It follows that $\|v'\|_1 = \|v\|_1 - \sum_{i=1}^{\tau} |v(q_i)|$.

For the independent Bernoulli approach, N_i is independent for each $i \in B$ and follows the Bernoulli distribution as

$$N_i = \begin{cases} 0 & \text{w.p.} \quad 1 - \frac{|v(i)|}{\|v'\|_1/(m-\tau)}\,, \\ 1 & \text{w.p.} \quad \frac{|v(i)|}{\|v'\|_1/(m-\tau)}\,. \end{cases}$$

Note that the probability is well defined due to the choice of τ. The number of nonzero components of the compressed vector is $\|\Phi_m(v)\|_0 = \tau + \sum_{i \in B} N_i$. From the choice of N_i, $\mathbb{E}\,(\|\Phi_m(v)\|_0 \mid v) = m$; so m is the expected sparsity of $\Phi_m(v)$.

Another choice is the systematic sampling approach [19]: Take a random variable U uniformly distributed in $(0, 1)$. Then for $k = 1, 2, \dots, m - \tau$, define

$$U_k = \frac{U + k - 1}{m - \tau}.$$

Given $\{q'_1, q'_2, \ldots, q'_{N-\tau}\}$ any permutation of indices in B, define

$$I_k = \max_{1\le i\le N-\tau}\left\{ i : \sum_{j=1}^{i-1} |v(q'_i)| \le U_k \|v'\|_1 < \sum_{j=1}^{i} |v(q'_i)| \right\}.$$

Then N_i is given by

$$N_i = \begin{cases} 1 & \text{if } i = q'_{I_k} \text{ for some } k, \\ 0 & \text{otherwise.} \end{cases}$$

Notice that by construction the number of nonzero N_i's is exactly $m-\tau$, and therefore $\|\Phi_m(v)\|_0 = m$. The N_i's generated by systematic sampling are obviously correlated as only one random number U drives the generation. The two approaches will be analyzed in the next section in the framework of the inexact power iteration.

3.2.2. Convergence analysis.

We now apply Theorem 1 to analyze the convergence of the FRI algorithm with either independent Bernoulli or systematic sampling. Notice that we have the following immediate result.

LEMMA 3. *Assumption* 3 *implies Assumptions* 1 *and* 2*.*

Therefore it suffices to check Assumption 3 for the compression function Φ_m. Homogeneity is obvious. From the construction of Φ_m, the unbiasedness is guaranteed by the expectation of N_i's, no matter which particular distribution is used for N_i.

$$\mathbb{E}\left(\Phi_m(v) \mid v\right) = v. \tag{37}$$

The variance bounds are proved in the following lemmas.

LEMMA 4. *For FRI compression with either independent Bernoulli or systematic sampling,*

$$\mathbb{E}\left(\|\Phi_m(v) - v\|_2^2 \mid v\right) \le \frac{\|v'\|_1^2}{m-\tau} \le \frac{\|v\|_1^2}{m}.$$

Moreover, we have the almost sure bound for systematic sampling,

$$\|\Phi_m(v) - v\|_2^2 \le \frac{2\|v'\|_1^2}{m-\tau} \le \frac{2\|v\|_1^2}{m} \quad a.s.$$

It is not possible to obtain an almost sure bound as above for independent Bernoulli, since, for example, it is possible that all the Bernoulli variables are 1, which gives the large error $\|\Phi_m(v) - v\|_2^2 \ge \frac{(N-2m+\tau)\|v'\|_1^2}{(m-\tau)^2}$. This lemma thus implies the advantage of using the systematic sampling strategy, which in practice gives smaller variance in general. We will only show numerical results using the systematic sampling strategy in the numerical examples later.

Proof. Since large components of v are remain unchanged by $\Phi_m(\cdot)$, we have

$$\|\Phi_m(v)-v\|_2^2 = \sum_{i=\tau+1}^{N} \left(\Phi_m(v)(q_i) - v(q_i)\right)^2 = \sum_{i=\tau+1}^{N} \big((\Phi_m(v)(q_i))^2 + v(q_i)^2 - 2\Phi_m(v)(q_i)v(q_i)\big).$$

Take the expectation

$$\begin{aligned}\mathbb{E}\left(\|\Phi_m(v)-v\|_2^2 \mid v\right) = \mathbb{E}&\left(\sum_{i=\tau+1}^N (\Phi_m(v)(q_i))^2 \mid v\right)\\ &+\sum_{i=\tau+1}^N v(q_i)^2 - 2\sum_{i=\tau+1}^N v(q_i)\mathbb{E}\left(\Phi_m(v)(q_i)\mid v\right).\end{aligned}$$

Since both independent Bernoulli and systematic sampling are unbiased,

$$\mathbb{E}\,\Phi_m(v)(q_i) = v(q_i).$$

Moreover, because there are $\sum_{i=\tau+1}^N N_i$ number of $\frac{\|v'\|_1}{m-\tau}$ and $n-\tau-\sum_{i=\tau+1}^N N_i$ number of 0 in $\{|\Phi_m(v)(q_i)|\}_{i\in B}$, we have

$$\begin{aligned}\mathbb{E}\left(\sum_{i=\tau+1}^N (\Phi_m(v)(q_i))^2 \mid v\right) &= \mathbb{E}\left(\mathbb{E}\left(\sum_{i=\tau+1}^N (\Phi_m(v)(q_i))^2 \mid \sum_{i=\tau+1}^N N_i\right)\mid v\right)\\ &= \frac{\|v'\|_1^2}{(m-\tau)^2}\mathbb{E}\left(\sum_{i=\tau+1}^N N_i \mid v\right).\end{aligned}$$

For independent Bernoulli, $\mathbb{E}\left(\sum_{i=\tau+1}^N N_i \mid v\right) = m-\tau$, and for systematic sampling, $\sum_{i=\tau+1}^N N_i = m-\tau$, so

$$\mathbb{E}\left(\sum_{i=\tau+1}^N (\Phi_m(v)(q_i))^2 \mid v\right) = \frac{\|v'\|_1^2}{m-\tau}.$$

Finally,

$$\mathbb{E}\left(\|\Phi_m(v)-v\|_2^2 \mid v\right) = \frac{\|v'\|_1^2}{m-\tau} - \|v'\|_2^2 \le \frac{\|v'\|_1^2}{m-\tau}.$$

We now show that $\frac{\|v'\|_1}{m-\tau} \le \frac{\|v\|_1}{m}$, which follows from the fact that $\frac{\sum_{j=i}^N |v(q_j)|}{m+1-i}$ is nonincreasing in i for $i \le \tau$. Indeed, recall from the choice of τ that for $i\le\tau$, $|v(q_i)| \ge \frac{\sum_{j=i}^N |v(q_j)|}{m+1-i}$, which is equivalent to

$$\frac{\sum_{j=i}^N |v(q_j)|}{m+1-i} \ge \frac{\sum_{j=i+1}^N |v(q_j)|}{m-i}.$$

Thus, combined with $\|v'\|_1 \le \|v\|_1$, we arrive at

$$\mathbb{E}\left(\|\Phi_m(v)-v\|_2^2 \mid v\right) \le \frac{\|v'\|_1^2}{m-\tau} \le \frac{\|v\|_1^2}{m}.$$

Next we give the almost sure bound for systematic sampling. Note that if $N_i \ne 0$ for $i\in B$, since $\big(\Phi_m(v)\big)(q_i)$ and $v(q_i)$ have the same sign, we have

$$(\Phi_m(v)(q_i) - v(q_i))^2 \le \Phi_m(q_i)^2 = \frac{\|v'\|_1^2}{(m-\tau)^2}.$$

Since there are exactly $m-\tau$ nonzero N_i's, we can estimate

$$\begin{aligned}\|\Phi_m(v)-v\|_2^2 &= \sum_{i=\tau+1}^{N}\big(\big(\Phi_m(v)\big)(q_i)-v(q_i)\big)^2\\ &\le \sum_{i=\tau+1}^{N} v(q_i)^2\mathbf{1}_{N_i=0}+\big(\Phi_m(v)\big)(q_i)^2\mathbf{1}_{N_i\neq 0}\\ &\le \|v'\|_2^2+(m-\tau)\frac{\|v'\|_1^2}{(m-\tau)^2}\\ &\le \frac{\|v'\|_1^2}{m-\tau}+\frac{\|v'\|_1^2}{m-\tau}=\frac{2\|v'\|_1^2}{m-\tau}\le\frac{2\|v\|_1^2}{m}. \qquad\square\end{aligned}$$

The expectation 1-norm bound can be easily checked from the definition.

Lemma 5. *For FRI with independent Bernoulli compression,*

$$\mathbb{E}\left(\|\Phi_m(v)\|_1 \mid v\right)=\|v\|_1.$$

For FRI with systematic sampling compression,

$$\|\Phi_m(v)\|_1=\|v\|_1 \quad a.s.$$

Therefore, the compression function Φ_m satisfies Assumption 3, and thus the convergence follows Theorem 1.

3.2.3. Deterministic compression by hard thresholding. Another way to choose the compression function Φ_m is by simple hard thresholding (HT), which means $\Phi_m=\Phi_m^{\mathrm{HT}}$ keeps the m largest entries (in absolute value) and drops the remaining ones. Compared to the previously discussed approaches of compression, HT obviously has smaller variance since it is deterministic, but pays a price for this by introducing bias to the inexact matrix-vector multiplication. The bias-variance trade-off between HT- and FRI-type algorithms is similar to that between iFCIQMC and FCIQMC.

4. Numerical results.

In this section, we give some numerical tests of the FCIQMC and FRI algorithms, and their variance iFCIQMC and HT to compare their performance. The numerical problem is to compute the ground state energy of a Hamiltonian H for a quantum system. As discussed before, we define $A=I-\delta H$ for δ small, so the problem is equivalent to finding the largest eigenvalue of A. We will test these methods with two types of model systems: the 2D fermionic Hubbard model and small chemical molecules under the full configuration interaction discretization. The Hamiltonians for these have the same structure. Each electron lives in a finite-dimensional one-particle Hilbert space. The vectors in the basis set of the one-particle Hilbert space are called orbitals. The number of orbitals N^{orb} is the dimension of the one-particle space. We denote by N^{elec} the total number of electrons in the system. Due to the Pauli exclusion principle, there are at most two electrons with opposite spins in one orbital. In our test examples, we choose the total spin $S^{\mathrm{tot}}=0$. Therefore the dimension of the space is $\binom{N^{\mathrm{orb}}}{N^{\mathrm{elec}}/2}^2$, neglecting other constraints like symmetry. The dimension grows exponentially as N^{orb} and N^{elec} grow. Here we summarize the system in our numerical tests in Table 1.

The exact ground state energy of the Hubbard model and Ne are computed using exact power iteration, and the ground state energy of H_2O is from the paper [22]. We

Table 1
Test systems.

System	N^{orb}	N^{elec}	Dimension	HF energy	Ground state energy
4×4 Hubbard	16	10	1.2×10^6	-17.7500	-19.5809
Ne, aug-cc-pVDZ	23	10	1.4×10^8	-128.4963	-128.7114
H_2O, cc-pVDZ	24	10	4.5×10^8	-76.0240	-76.2419

use HANDE-QMC[29] (http://www.hande.org.uk/), an open source stochastic quantum chemistry program written in Fortran, for FCIQMC and iFCIQMC calculation. FRI and HT subroutines are implemented in Fortran based on HANDE-QMC. The Hamiltonian of the Hubbard model is included in the HANDE-QMC package, and the Hamiltonians of Ne and H_2O are calculated using RHF (restricted Hartree–Fock) by Psi4 (http://www.psicode.org/), an open source ab initio electronic structure package. The code to generate the entries of Hamiltonian H is the same for all four algorithms, so the comparison among algorithms is fair in terms of computational time. The four algorithms are tested on a computer with 6 Core Xeon CPU at 3.5GHz and 64GB RAM.

Note that our comparison is mostly for illustrative purpose and should not be taken as benchmark tests for the various algorithms, especially for large-scale calculations, which would depend on parallel implementation, hardware infrastructure, etc. On the other hand, even for small problems, the numerical results still offer some suggestions on further development of inexact power iteration–based solvers for many-body quantum systems.

4.1. Hubbard model. The Hubbard model is a standard model used in condensed matter physics, which describes interacting particles on a lattice. In real space, the Hubbard Hamiltonian is

$$H = -\sum_{\langle r,r'\rangle,\sigma} \hat{c}^\dagger_{r,\sigma}\hat{c}_{r',\sigma} + U\sum_r \hat{n}_{r\uparrow}\hat{n}_{r\downarrow}, \tag{38}$$

where we have scaled the hopping parameter to be 1, and so the on-site repulsion parameter U gives the ratio of interaction strength relative to the kinetic energy. We choose an intermediate interaction strength $U = 4$ in our test.

In the d-dimensional Hubbard Hamiltonian (38), r is a d-dimensional vector representing a site in the lattice, $\langle r, r'\rangle$ means r and r' are nearest neighbors to each other, and σ takes values of $\uparrow$ and $\downarrow$, which is the spin of the electron. $\hat{c}_{r,\sigma}$ and $\hat{c}^\dagger_{r,\sigma}$ are the annihilation and creation operator of electrons at site r with spin σ. They satisfy the commutation relations

$$\{\hat{c}_{r,\sigma}, \hat{c}^\dagger_{r',\sigma'}\} = \delta_{r,r'}\delta_{\sigma,\sigma'}, \qquad \{\hat{c}_{r,\sigma}, \hat{c}_{r',\sigma'}\} = 0, \quad \text{and} \quad \{\hat{c}^\dagger_{r,\sigma}, \hat{c}^\dagger_{r',\sigma'}\} = 0,$$

where $\{A, B\} = AB + BA$ is the anticommutator. $\hat{n}_{r,\sigma}$ is the number operator and defined as $\hat{n}_{r,\sigma} = \hat{c}^\dagger_{r,\sigma}\hat{c}_{r,\sigma}$. We will consider the Hubbard model on a finite 2D lattice with periodic boundary condition.

When the interaction strength U is small, it is better to work in the momentum space instead of the real space, since the planewaves are the eigenfunctions of the kinetic part of the Hamiltonian. The annihilation operator in momentum space is $\hat{c}_{k,\sigma} = \frac{1}{\sqrt{N^{\text{orb}}}}\sum_r e^{ik\cdot r}\hat{c}_{r,\sigma}$, where $k = (k_1, k_2)$ is the wave number and N^{orb} is the total number of orbitals or sites. The Hubbard Hamiltonian in momentum space is then

$$H = \sum_{k,\sigma} \varepsilon(k) n_{k,\sigma} + \frac{U}{N^{\mathrm{orb}}} \sum_{k,p,q} c^{\dagger}_{p-q,\uparrow} c^{\dagger}_{k+q,\downarrow} c_{k,\downarrow} c_{p,\uparrow}, \tag{39}$$

where $\varepsilon(k) = -2\sum_{i=1}^{2}\cos(k_i)$.

Written as a matrix, the Hubbard Hamiltonian in the momentum space is just a real symmetric matrix with diagonal entries $\varepsilon(k)$ and off-diagonal either 0 or $\pm\frac{U}{N^{\mathrm{orb}}}$. For inexact power iteration, we take $A = I - \delta H$ with $\delta = 0.01$. In our numerical test, we will use the projected energy estimator for the smallest eigenvalue of H; the projected vector v_* is chosen to be the Hartree–Fock state. The initial iteration of all methods is also chosen as the Hartree–Fock state (a vector whose only nonzero entry is at the Slater determinant corresponding to the Hartree–Fock ground state of the system).

Figure 1 plots the error of projected energy of each iteration versus wall-clock time (first 1500 seconds) for a typical realization. The error is defined as the difference between the projected energy estimate and the exact ground state energy. The complexity parameters of the algorithms are shown in Table 2, which are chosen such that FRI and FCIQMC use about the same amount of memory (e.g., the particle number in FCIQMC is roughly equal to the nonzero entries of the matrix-vector product in FRI or HT before compression), and also chosen so large that all the algorithms converge. The time per iteration listed in Table 2 is averaged over several realizations and is used in Figure 1.

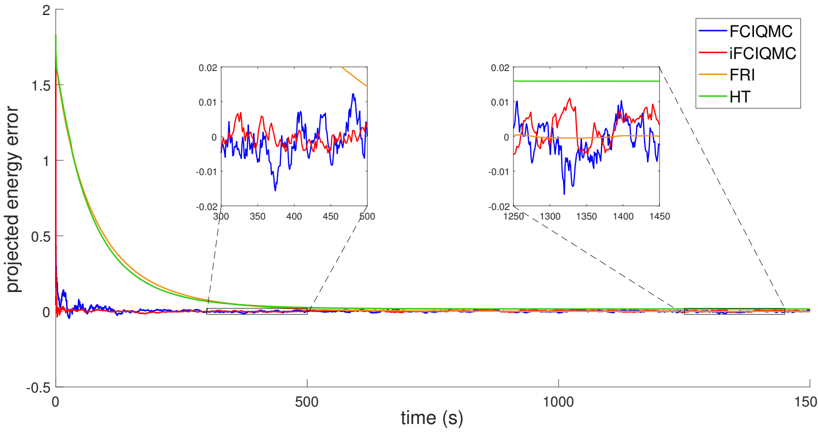

Fig. 1. *Convergence of the projected energy with respect to time for system* 1*, a* 4×4 *Hubbard model with* 10 *electrons,* 5 *spin up and* 5 *spin down, and interaction strength* $U = 4$*. (See color online.)*

As shown in Figure 1, all four algorithms converge to result close to the exact eigenvalue, and the estimated value from each iteration stays around the eigenvalue for a long time. FCIQMC and *i*FCIQMC take much less time to converge, thanks to their much lower cost inexact matrix-vector multiplication compared to FRI and HT, but the variance is also larger. In terms of iteration number, the convergence of the four algorithms is similar, which can be understood from our analysis since it is the same eigenvalue gap of the Hamiltonian that drives the convergence. As

Table 2

Parameters and numerical results for 4×4 *Hubbard model with* 10 *electrons,* 5 *spin up and* 5 *spin down, and interaction strength* $U = 4$*.*

	m	$\|Av_t\|_0$	Avg. error	Std.	MSE	τ_{auto}	Compr. error	Time/iter.(s)
FCIQMC	1.7×10^6	-	4.4×10^{-4}	3.0×10^{-4}	2.6×10^{-7}	14.1	4.6×10^{-2}	1.1
iFCIQMC	1.7×10^6	-	3.2×10^{-4}	2.2×10^{-4}	1.8×10^{-7}	13.8	2.6×10^{-2}	0.91
FRI	3.0×10^4	9.4×10^5	1.2×10^{-4}	6.1×10^{-5}	2.8×10^{-8}	13.7	1.6×10^{-1}	3.6
HT	3.0×10^4	7.2×10^5	1.6×10^{-2}	-	2.5×10^{-4}	-	4.5×10^{-3}	3.5

we mentioned already, per iteration, the FCIQMC and iFCIQMC is much cheaper in comparison. The reason is that FRI and HT need to access all nonzero elements of A for each column associated with a nonzero entry in the current iterate (for multiplying A with the sparse vector), while FCIQMC and iFCIQMC just need to randomly pick some without accessing the others. The number of nonzero entries per row is large, and accessing elements of A is quite expensive for FCI-type problems. More quantitatively, we see in Table 2 that for a sparse vector of 3×10^4 nonzero entries in FRI, after multiplication by A before compression, the number of nonzero entries increases to roughly 10^6. Thus for this problem, on average, each column has about 40 nonzero entries that FRI needs to access, while the FCIQMC algorithm only needs access to a few entries after the random choice.

After convergence, the projected energies of FCIQMC and iFCIQMC fluctuate around the exact ground state energy. Although iFCIQMC is biased, the bias is not large for the current problem, while the variance is smaller than that of FCIQMC. So iFCIQMC is an effective strategy for bias-variance trade-off. The projected energy of FRI also varies around the true energy, and the variance is much smaller than that of FCIQMC or iFCIQMC. HT is deterministic, and the projected energy shows no variance. However, the bias is also quite visible.

We can average the projected energy over the path to get a better estimate. The variance of the estimator will decay to zero as we include a longer time period in the average. Thus, due to unbiasedness, the error of FCIQMC and FRI can be made smaller if we run them long enough. In Table 2, we give more quantitative comparison of the results of the algorithms. The quantities in the table are defined as below:

Avg. error
$$\frac{1}{w}\sum_{i=i_0}^{i_0+w-1} |E_i - E^{\text{true}}|$$

Std.
$$\sqrt{\frac{1}{w-1}\sum_{i=i_0}^{i_0+w-1}\Big(E_i - \frac{1}{w}\sum_{j=i_0}^{i_0+w-1} E_i\Big)^2}\sqrt{\frac{1+2\tau_{\text{auto}}}{W}}$$

MSE
$$\text{avg. error}^2 + \text{std.}^2$$

τ_{auto}
$$\sum_{t=1}^{w-1} \frac{\frac{1}{w-1}\sum_{i=i_0}^{i_0+w-t-1}\Big(E_i - \frac{1}{w}\sum_{j=i_0}^{i_0+w-1} E_i\Big)\Big(E_{i+t} - \frac{1}{w}\sum_{j=i_0}^{i_0+w-1} E_i\Big)}{\frac{1}{w-1}\sum_{i=i_0}^{i_0+w-1}\Big(E_i - \frac{1}{w}\sum_{j=i_0}^{i_0+w-1} E_i\Big)^2}$$

Compr. error
$$\frac{1}{w}\sum_{i=i_0}^{i_0+w-1} \frac{\|\xi_{t+1}\|_2}{\|Av_t\|_2}$$

Here E^{true} is the true ground state energy obtained by exact power iteration, i_0 is a burn-in parameter, and w is the window size of the average. For FCIQMC and

iFCIQMC, $w = 1600$ and $i_0 = 2400$. For FRI and HT, $w = 400$ and $i_0 = 600$. The numerical tests show that the quantities above are insensitive to the choice of w and i_0, as long as the algorithms indeed converge after i_0 steps and the window size w is not too small. τ_{auto} is the integrated autocorrelation time, and W is the number of iterations averaged. Std. is short for the standard deviation of the sample mean $\bar{E}^{(W)}$ defined as $\bar{E}^{(W)} = \frac{1}{W}\sum_{i=i_0}^{i_0+W-1} E_i$. Since the time cost per iteration of different algorithms is quite different, to make a fair comparison, we take $W = \frac{10000}{\text{time per iter.}}$ for each algorithm. This gives the standard error of the sample mean if we run each algorithm for 10000 seconds after convergence. The mean square error (MSE) is simply defined to incorporate the variance and bias together.

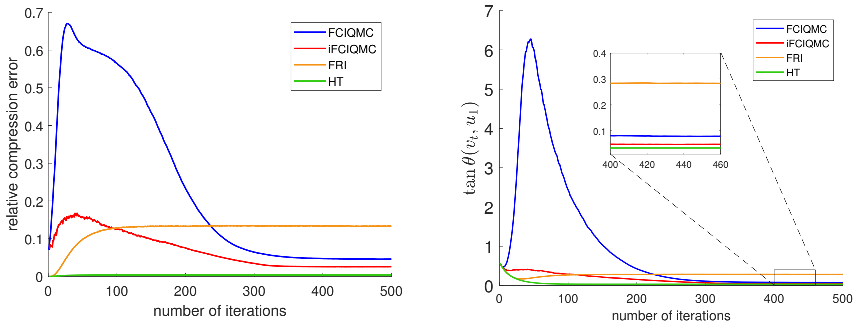

Fig. 2. *4×4 Hubbard model. Left: Relative compression error $\|\xi_{t+1}\|_2/\|Av_t\|_2$ as a function of iteration steps. Right: Angle between the iterate and the exact ground state* $\tan\theta(v_t, u_1)$. *(See color online.)*

To obtain further insight into the interplay between the error per step of the inexact power iteration and the convergence, we plot in Figure 2 the relative compression error and the tangent of the angle between v_t and the exact eigenvector u_1. We observe that FRI and HT reach convergence after about 100 steps, and FCIQMC and iFCIQMC converge after about 350 steps; the greater number of steps needed by FCIQMC and iFCIQMC is related to the first phase of the algorithm where the particle number is exponentially growing. This can be seen from Figure 2 (left) as the huge error growth of the initial stage of the iterations. Only when the particle number reaches a certain level does the compression error become small and the power iteration convergence kick in.

After convergence, FRI has the largest compression error and HT has the smallest. The compression error of iFCIQMC is also smaller than that of FCIQMC. It is reasonable since HT and iFCIQMC reduce variance and thus compression error compared with the fully stochastic FRI and FCIQMC. As shown in Figure 2, in this example with the parameter choice, FCIQMC has a smaller compression error than FRI, and the larger the compression error is, the further v_t is away from the true eigenvector u_1. This agrees with the theoretical results we obtain in Theorem 1, because $\tan\theta(v_t, u_1)$ is controlled by the error ξ_t at each step.

We remark that the $\tan\theta(v_t, u_1)$ error measure does not directly translate to the error of the projected energy estimator using, say, the Hartree–Fock state. In fact, we observe in Figure 1 and Table 2 that, per iteration, the projected energy estimated by FRI is smaller than FCIQMC and iFCIQMC. As an explanation, in our parameter regime, the exact ground state has a large overlap with the Hartree–Fock state, so in FRI, that component is kept unchanged in the compression, while for FCIQMC

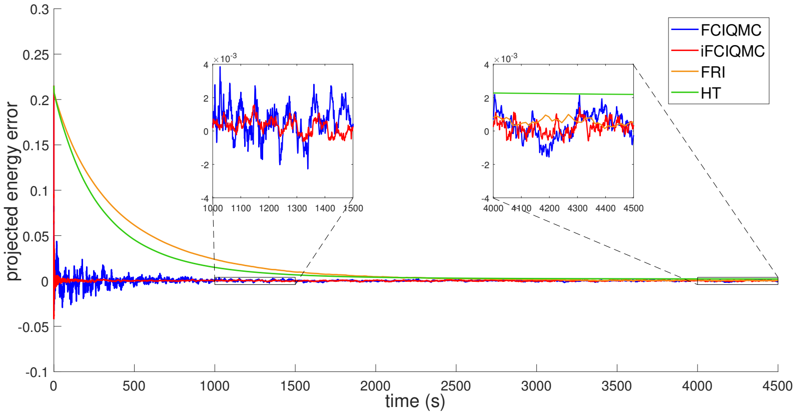

FIG. 3. *Convergence of the projected energy with respect to time for Ne in aug-cc-pVDZ basis. (See color online.)*

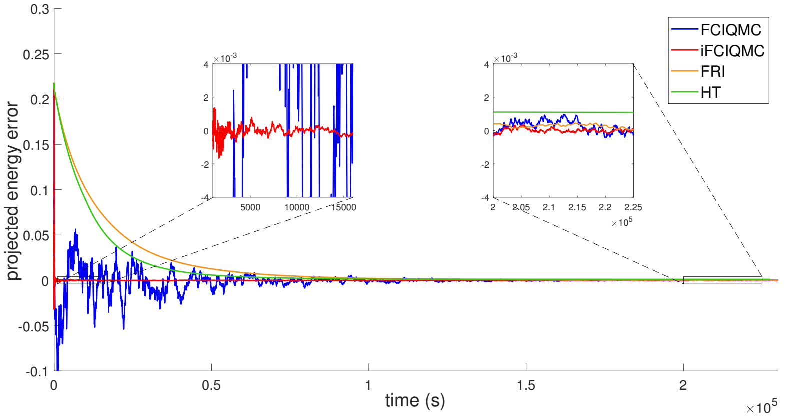

FIG. 4. *Convergence of the projected energy with respect to time for H_2O in cc-pVDZ basis. (See color online.)*

and iFCIQMC, the stochastic error is more uniformly distributed over all the entries. This behavior seems more problem dependent though, and we will see in the chemical molecular examples that the MSE of FRI becomes comparable to that of FCIQMC.

4.2. Molecules. We also tested the four algorithms for some molecule examples. The FCI Hamiltonian is obtained by Hartree–Fock calculations in a chosen chemical basis (for single-particle Hilbert space), such as cc-PVDZ. We choose Ne and H_2O at equilibrium geometry as examples, which is described in Table 1. The time step is taken as $\delta = 0.01$.

The convergence of projected energy error versus wall-clock time is shown in Figures 3 and 4, respectively. The parameter choice of the algorithms and more quantitative comparison are shown in Tables 3 and 4. The four algorithms also work well for molecule systems. The convergence behavior is similar to the Hubbard case.

The complexity parameter m needed to achieve convergence depends on the sys-

Table 3
Comparison of algorithms for Ne in aug-cc-pVDZ basis.

	m	$\|Av_t\|_0$	Avg. error	Std.	MSE	τ_{auto}	Time/iter.(s)
FCIQMC	1.8×10^6	-	1.1×10^{-4}	3.8×10^{-5}	1.8×10^{-8}	12.4	1.5
iFCIQMC	1.8×10^6	-	7.7×10^{-5}	2.4×10^{-5}	9.6×10^{-9}	15.3	1.1
FRI	1.0×10^4	7.9×10^6	8.0×10^{-5}	4.8×10^{-5}	1.1×10^{-8}	12.3	11.6
HT	1.0×10^4	3.3×10^6	1.7×10^{-3}	-	2.8×10^{-6}	-	7.9

Table 4
Comparison of algorithms for H_2O in cc-pVDZ basis.

	m	$\|Av_t\|_0$	Avg. error	Std.	MSE	τ_{auto}	Time/iter.(s)
FCIQMC	6.0×10^7	-	4.1×10^{-5}	1.7×10^{-4}	2.1×10^{-9}	25.6	54.1
iFCIQMC	6.0×10^7	-	1.4×10^{-5}	5.3×10^{-5}	2.7×10^{-10}	119	36.6
FRI	1.2×10^5	1.6×10^8	2.2×10^{-5}	1.2×10^{-4}	8.9×10^{-10}	12.8	379.3
HT	1.2×10^5	3.4×10^7	1.1×10^{-3}	-	1.2×10^{-6}	-	227.0

tem. The ratio m/N of Ne is smaller than H_2O. The time cost of FRI and HT is much larger than FCIQMC and iFCIQMC, because they require the exact matrix-vector multiplication Av_t, which is still expensive although v_t is sparse. Unlike the Hubbard case where FRI gives much smaller error, the MSE of FRI is similar to FCIQMC and iFCIQMC in these cases.

In summary, the numerical examples show that FCIQMC, FRI, and their variants can achieve convergence using much less memory and computational time compared to the standard power iteration. The stochastic algorithms FCIQMC, iFCIQMC, and FRI give better estimates than the deterministic method HT in general. The numerical test also points out directions to further improve these inexact power iterations, including variance and memory cost reduction of the inexact matrix-vector multiplication and efficient parallel implementation to overcome the memory bottleneck. These will be left for future work.

Acknowledgments. We are grateful for useful discussions with George Booth, Yingzhou Li, Jonathan Weare, Stephen Wright, and Lexing Ying during various stages of this work.

REFERENCES

[1] M. L. Abrams and C. David Sherrill, *Important configurations in configuration interaction and coupled-cluster wave functions*, Chem. Phys. Lett., 412 (2005), pp. 121–124.

[2] N. S. Blunt, S. D. Smart, J. A. F. Kersten, J. S. Spencer, G. H. Booth, and A. Alavi, *Semi-stochastic full configuration interaction quantum Monte Carlo: Developments and application*, J. Chem. Phys., 142 (2015), 184107, http://aip.scitation.org/doi/10.1063/1.4920975.

[3] G. H. Booth, D. Cleland, A. J. W. Thom, and A. Alavi, *Breaking the carbon dimer: The challenges of multiple bond dissociation with full configuration interaction quantum Monte Carlo methods*, J. Chem. Phys., 135 (2011), 084104, http://aip.scitation.org/doi/10.1063/1.3624383.

[4] G. H. Booth, A. Grüneis, G. Kresse, and A. Alavi, *Towards an exact description of electronic wavefunctions in real solids*, Nature, 493 (2013), pp. 365–370.

[5] G. H. Booth, A. J. W. Thom, and A. Alavi, *Fermion Month Carlo without fixed nodes: A game of life, death, and annihilation in Slater determinant space*, J. Chem. Phys., 131 (2009), 054106.

[6] R. J. Buenker and S. D. Peyerimhoff, *Individualized configuration selection in CI calculations with subsequent energy extrapolation*, Theoretica Chimica Acta, 35 (1974), pp. 33–58, http://link.springer.com/10.1007/BF02394557.

[7] D. Cleland, G. H. Booth, and A. Alavi, *Survival of the fittest: Accelerating convergence in full configuration-interaction quantum Monte Carlo*, J. Chem. Phys., 132 (2010), 041103.

[8] J.-P. Daudey, J.-L. Heully, and J.-P. Malrieu, *Size-consistent self-consistent truncated or selected configuration interaction*, J. Chem. Phys., 99 (1993), pp. 1240–1254, http://aip.scitation.org/doi/10.1063/1.465368.

[9] F. A. Evangelista, *Adaptive multiconfigurational wave functions*, J. Chem. Phys., 140 (2014), 124114, http://aip.scitation.org/doi/10.1063/1.4869192.

[10] E. Giner, A. Scemama, and M. Caffarel, *Using perturbatively selected configuration interaction in quantum Monte Carlo calculations*, Canad. J. Chem., 91 (2013), pp. 879–885, http://www.nrcresearchpress.com/doi/abs/10.1139/cjc-2013-0017.

[11] J. C. Greer, *Estimating full configuration interaction limits from a Monte Carlo selection of the expansion space*, J. Chem. Phys., 103 (1995), pp. 1821–1828, http://aip.scitation.org/doi/10.1063/1.469756.

[12] J. C. Greer, *Monte Carlo configuration interaction*, J. Comput. Phys., 146 (1998), pp. 181–202, http://www.sciencedirect.com/science/article/pii/S0021999198959538.

[13] R. J. Harrison, *Approximating full configuration interaction with selected configuration interaction and perturbation theory*, J. Chem. Phys., 94 (1991), pp. 5021–5031, http://aip.scitation.org/doi/10.1063/1.460537.

[14] A. A. Holmes, H. J. Changlani, and C. J. Umrigar, *Efficient heat-bath sampling in Fock space*, J. Chem. Theory Comput., 12 (2016), pp. 1561–1571.

[15] B. Huron, J. P. Malrieu, and P. Rancurel, *Iterative perturbation calculations of ground and excited state energies from multiconfigurational zeroth-order wavefunctions*, J. Chem. Phys., 58 (1973), pp. 5745–5759, http://aip.scitation.org/doi/10.1063/1.1679199.

[16] F. Illas, J. Rubio, J. M. Ricart, and P. S. Bagus, *Selected versus complete configuration interaction expansions*, J. Chem. Phys., 95 (1991), pp. 1877–1883, http://aip.scitation.org/doi/10.1063/1.461037.

[17] J. Ivanic, *Direct configuration interaction and multiconfigurational self-consistent-field method for multiple active spaces with variable occupations.* I. *Method*, J. Chem. Phys., 119 (2003), pp. 9364–9376, http://aip.scitation.org/doi/10.1063/1.1615954.

[18] P. J. Knowles, *Compressive sampling in configuration interaction wavefunctions*, Molecular Phys., 113 (2015), pp. 1655–1660, https://www.tandfonline.com/doi/full/10.1080/00268976.2014.1003621.

[19] L.-H. Lim and J. Weare, *Fast randomized iteration: Diffusion Monte Carlo through the lens of numerical linear algebra*, SIAM Rev., 59 (2017), pp. 547–587, https://doi.org/10.1137/15M1040827.

[20] W. Liu and M. R. Hoffmann, *iCI: Iterative CI toward full CI*, J. Chem. Theory Comput., 12 (2016), pp. 1169–1178, http://pubs.acs.org/doi/abs/10.1021/acs.jctc.5b01099.

[21] D. Ma, G. Li Manni, and L. Gagliardi, *The generalized active space concept in multiconfigurational self-consistent field methods*, J. Chem. Phys., 135 (2011), 044128, http://aip.scitation.org/doi/10.1063/1.3611401.

[22] J. Olsen, *Full configuration-interaction and state of the art correlation calculations on water in a valence double-zeta basis with polarization functions*, J. Chem. Phys., 104 (1996), pp. 8007–8015, http://aip.scitation.org/doi/10.1063/1.471518.

[23] F. R. Petruzielo, A. A. Holmes, H. J. Changlani, M. P. Nightingale, and C. J. Umrigar, *Semistochastic projector Monte Carlo method*, Phys. Rev. Lett., 109 (2012), 230201.

[24] Z. Rolik, Á. Szabados, and P. R. Surján, *A sparse matrix based full-configuration interaction algorithm*, J. Chem. Phys., 128 (2008), 144101, http://aip.scitation.org/doi/10.1063/1.2839304.

[25] R. Roth, *Importance truncation for large-scale configuration interaction approaches*, Phys. Rev. C, 79 (2009), 064324, https://link.aps.org/doi/10.1103/PhysRevC.79.064324.

[26] J. B. Schriber and F. A. Evangelista, *Communication: An adaptive configuration interaction approach for strongly correlated electrons with tunable accuracy*, J. Chem. Phys., 144 (2016), 161106, http://aip.scitation.org/doi/10.1063/1.4948308.

[27] L. R. Schwarz, A. Alavi, and G. H. Booth, *Projector quantum Monte Carlo method for nonlinear wave functions*, Phys. Rev. Lett., 118 (2017), 176403, http://link.aps.org/doi/10.1103/PhysRevLett.118.176403.

[28] S. Sharma, A. A. Holmes, G. Jeanmairet, A. Alavi, and C. J. Umrigar, *Semistochastic heat-bath configuration interaction method: Selected configuration interaction with semistochastic perturbation theory*, J. Chem. Theory Comput., 13 (2017), pp. 1595–1604, http://pubs.acs.org/doi/abs/10.1021/acs.jctc.6b01028.

[29] J. S. Spencer, N. S. Blunt, W. A. Vigor, F. D. Malone, W. M. C. Foulkes, J. J. Shepherd, and A. J. W. Thom, *Open-source development experiences in scientific soft-*

ware: The HANDE Quantum Monte Carlo Project, J. Open Res. Software, 3 (2015), https://doi.org/10.5334/jors.bw.

[30] S. Ten-no, *Stochastic determination of effective Hamiltonian for the full configuration interaction solution of quasi-degenerate electronic states*, J. Chem. Phys., 138 (2013), 164126, http://aip.scitation.org/doi/10.1063/1.4802766.

[31] N. M. Tubman, J. Lee, T. Y. Takeshita, M. Head-Gordon, and K. B. Whaley, *A deterministic alternative to the full configuration interaction quantum Monte Carlo method*, J. Chem. Phys., 145 (2016), 044112.

[32] T. Zhang and F. A. Evangelista, *A deterministic projector configuration interaction approach for the ground state of quantum many-body systems*, J. Chem. Theory Comput., 12 (2016), pp. 4326–4337, http://pubs.acs.org/doi/abs/10.1021/acs.jctc.6b00639.

[33] P. M. Zimmerman, *Incremental full configuration interaction*, J. Chem. Phys., 146 (2017), 104102, http://aip.scitation.org/doi/10.1063/1.4977727.

[34] P. M. Zimmerman, *Strong correlation in incremental full configuration interaction*, J. Chem. Phys., 146 (2017), 224104, http://aip.scitation.org/doi/10.1063/1.4985566.