

Unbiased Reduced Density Matrices and Electronic Properties from Full Configuration Interaction Quantum Monte Carlo

Catherine Overy,¹ George H. Booth,^{1,2,3, a)} N. S. Blunt,¹ James J. Shepherd,¹ Deidre Cleland,^{1,4} and Ali Alavi^{1,5, b)}

¹⁾ *University of Cambridge, Chemistry Department, Lensfield Road, Cambridge CB2 1EW, United Kingdom*

²⁾ *King's College London, Theory and Simulation of Condensed Matter, The Strand, London WC2R 2LS, United Kingdom*

³⁾ *Thomas Young Centre, University College London, 17 Gordon Street, London WC1H 0AH, United Kingdom*

⁴⁾ *CSIRO Virtual Nanoscience Laboratory, 343 Royal Parade, Parkville, Victoria 3052, Australia*

⁵⁾ *Max Planck Institute for Solid State Research, Heisenbergstr. 1, D-70569, Stuttgart, Germany*

(Dated: 10 May 2018)

Properties that are necessarily formulated within pure (symmetric) expectation values are difficult to calculate for projector quantum Monte Carlo approaches, but are critical in order to compute many of the important observable properties of electronic systems. Here, we investigate an approach for the sampling of unbiased reduced density matrices within the Full Configuration Interaction Quantum Monte Carlo dynamic, which requires only small computational overheads. This is achieved via an independent replica population of walkers in the dynamic, sampled alongside the original population. The resulting reduced density matrices are free from systematic error (beyond those present via constraints on the dynamic itself), and can be used to compute a variety of expectation values and properties, with rapid convergence to an exact limit. A quasi-variational energy estimate derived from these density matrices is proposed as an accurate alternative to the projected estimator for multiconfigurational wavefunctions, while its variational property could potentially lend itself to accurate extrapolation approaches in larger systems.

I. INTRODUCTION

The extraction of expectation values and properties of quantum systems whose operator does not commute with the Hamiltonian has been a significant hurdle in the widespread adoption of projector Quantum Monte Carlo methods for many years¹. These include important physical properties such as the dipole or higher electrical moments, as well as particle distribution functions, forces and higher derivatives of nuclei position, static correlation functions, quantum entropy estimators and various order parameters in the condensed phase. As such, their reliable and unbiased computation within the projector Quantum Monte Carlo (QMC) framework is an active and important research area^{2–9}. Within continuum Diffusion Monte Carlo and its variants, there exist techniques to unbiased these quantities such as ‘forward-walking’ and its variants^{3–6}, Reptation Quantum Monte Carlo⁷, and Hellmann–Feynman sampling^{8,9}, all of which have drawbacks regarding the additional computational effort they require.

Recently, the Full Configuration Interaction Quantum Monte Carlo method (FCIQMC) has been introduced as a projector QMC technique formulated within the second quantized algebra of arbitrary (antisymmetrized) Hilbert

spaces. This method relies on the fact that, as opposed to continuum approaches, the space is finite (although exponentially large) and therefore can employ annihilation (cancellation) processes between walkers of different signs in order to overcome the Fermion sign problem^{10–12}. Additional systematically improvable approximations have also been introduced to the method to allow for a large reduction in computational effort, while nevertheless converging in the limit of large sampling to exact results within the basis employed^{13–15}. Furthermore, there is significant empirical evidence that the computational effort required for the sampling of many systems scales sub-linearly with the size of the Hilbert space as systems grow in size and complexity^{14,16,17}. This has allowed for near-exact results to be obtained within systematically improvable random error bars with substantially less computational effort than an iterative diagonalization approach to the problem. This has allowed for large molecular¹⁸, solid state¹⁹ and model Hamiltonians^{12,20–22} to be investigated with the technique.

However, studies to date have primarily focused on properties of systems related to total energy differences, such as ionization potentials, electron affinities, potential energy surfaces and equations of state, all of which can be computed in an unbiased fashion through calculation of a *projected* energy expression. This relies on the fact that for an eigenstate $|\Psi\rangle$, with energy E , this energy

^{a)} Electronic mail: george.booth24@gmail.com

^{b)} Electronic mail: asa10@cam.ac.uk

can be reproduced from

$$E = \langle E_P \rangle = \frac{\langle \psi_T | \hat{H} | \Psi \rangle}{\langle \psi_T | \Psi \rangle}, \quad (1)$$

for any function $|\psi_T\rangle$ with a non-zero overlap with $|\Psi\rangle$. However, to extend the scope of the method, it is necessary to be able to extract unbiased properties from the sampled wavefunctions, for operators which do not commute with the Hamiltonian, such as those which are a function of position. In these cases, Eq. 1 will not be sufficient if the operator of interest has a different set of eigenstates⁴.

Formally, these properties can be derived from the presence of a perturbing operator to the Hamiltonian, \hat{A} , as

$$\hat{H}' = \hat{H} + \lambda \hat{A}, \quad (2)$$

where λ is the strength of the perturbation. The relevant expectation value associated with \hat{A} can then be obtained by calculating the analytical (or numerical) gradient of the energy with respect to the perturbation strength, at zero perturbation²³,

$$\langle A \rangle = \left. \frac{\partial E'}{\partial \lambda} \right|_{\lambda=0}. \quad (3)$$

In the case of stationary wavefunctions, which is the case (on average) in FCIQMC, the Hellmann–Feynman theorem²⁴ reduces this to the calculation of a *pure* expectation value of the operator, \hat{A} , as

$$\langle A \rangle = \frac{\langle \Psi | \hat{A} | \Psi \rangle}{\langle \Psi | \Psi \rangle}. \quad (4)$$

Equivalently, this can be written as the trace of \hat{A} with the appropriate rank reduced density matrix for the number of coupled particles in \hat{A} . In continuum approaches, the sampled distribution consists of the desired wavefunction multiplied by a trial wavefunction²⁵, whereas in FCIQMC, a single (pure) wavefunction distribution is sampled²⁶. However, this does not remove an inherent bias from these expectation values, which fundamentally arises from the difficulty in evaluating quadratic functionals of stochastically sampled wavefunctions², as found in Eq. 4. Strictly, a similar bias arises even in naive accumulations of the projected energy estimator of Eq. (1). This is because the ratio of the sampled wavefunction distributions has a finite covariance which needs to be properly accounted for when calculating the quantity^{27,28}.

In this paper, we describe the source of this bias for quadratic functionals, before demonstrating a way to sample one- and two-body reduced density matrices (RDMs) of the sampled wavefunction within the FCIQMC dynamic in an unbiased fashion, with only small additional computational overheads. This allows for calculation of *pure* expectation values and quantities of interest via a trace of the resultant N -representable reduced density matrices with the appropriate operator^{29–31}. The accumulation of the one- and

two-body RDMs therefore allows for access to one- and two-body expectation values of the wavefunction, assuming that the appropriate operator can be projected into the space.

Indeed, the energy is also a two-body expectation value (of the Hamiltonian), and therefore an alternative, variational estimate of the energy is also accessible from the two-body RDM. There may be significant advantages to an accurate computation of this quantity, despite the availability of the projected estimate (Eq. 1), since the pure estimate does not rely on a good quality trial wavefunction in order to obtain a large overlap with the true wavefunction and minimize stochastic fluctuations of the energy. This is likely to be important in multiconfigurational systems with significant static correlation, whilst the variationality of the estimator may also offer a more rigorous extrapolation procedure for systematic errors compared to the non-variational projected estimate. Furthermore, as a quadratic functional of the wavefunction, errors in the energy given by the pure expectation value of Eq. 4 will scale quadratically with the error in the wavefunction, rather than linearly in the case of Eq. 1, again potentially leading to an improved estimator.

Section II will briefly recap the stochastic FCIQMC algorithm, before detailing an adaptation to allow for non-integer walker weights to reduce the random error-bars in the approach. Section III will then detail a naive approach to sampling density matrices, which was implemented in Ref. 32, but which was known to be biased, the origin of which is pinpointed and discussed. An updated approach is then formulated and presented in section IV, which rigorously removes this bias, and the convergence and unbiased nature of the resulting properties are demonstrated. The computational overhead of the approach is analysed in detail, and strategies to minimize its impact without introducing further approximations illustrated. Finally, rigorous benchmarking of a range of properties are shown, including the quasi-variational energy estimate, all of which show encouraging accuracy and convergence properties with both imaginary time, and walker number.

II. FCIQMC RECAPITULATION

Full Configuration Interaction Quantum Monte Carlo can be considered as a stochastic minimization of the energy with respect to a sampled full configuration interaction wavefunction expansion. This wavefunction is a simple linear combination of all Slater determinants that can be constructed from distributing the available electrons within the (orthonormalized) single-particle orbitals spanning the space, as

$$|\Psi\rangle = \sum_{\mathbf{i}} C_{\mathbf{i}} |D_{\mathbf{i}}\rangle, \quad (5)$$

where $|D_{\mathbf{i}}\rangle$ represents a Slater determinant, labeled by the orbital occupation string \mathbf{i} . The linear coefficients

of this expansion are the objects that are stochastically sampled, such that the average sampled determinant weights $\langle n_i \rangle$ are proportional to C_i . Their optimization to the variational ground state of the ansatz is simulated via a stochastic, iterative application of the equations:

$$\Delta n_i(\beta + \tau) = -\tau \left[\sum_{j \neq i} H_{ij} n_j(\beta) \right] - \tau (H_{ii} - E_S) n_i(\beta) \quad (6)$$

where $\Delta n_i(\beta)$ represents the change in ‘walker’ population/weight on determinant $|D_i\rangle$ in the time step $\beta \rightarrow \beta + \tau$. This leads to population dynamics of a set of walkers which occupy determinants connected to each other in this many-electron Hilbert space. This dynamic consists of a set of stochastically realized processes.

The first is a ‘spawning’ step, which is performed for each occupied determinant, and a number of times proportional to the walker weight at that determinant (n_i). A single or double excitation is randomly chosen, with normalized probability $p_{\text{gen}}(\mathbf{j}|\mathbf{i})$ for the excitation from $|D_i\rangle$ to $|D_j\rangle$. The walker amplitude on $|D_j\rangle$ is then augmented with a signed probability given by

$$p_{\text{spawn}} = -\frac{\tau H_{ij}}{p_{\text{gen}}(\mathbf{j}|\mathbf{i})}. \quad (7)$$

Finally, a ‘death’ step is performed, by which the amplitude on each determinant, $|D_i\rangle$, is (generally) reduced with probability $\tau(H_{ii} - E_S)n_i$. Taken together, these two steps simulate the dynamic in Eq. 6. However, an additional ‘annihilation’ step is essential in order to overcome an exponential increase in noise and other features associated with the Fermion sign problem^{12,33}. In this step, walkers of opposite signs on the same determinant are removed from the simulation.

As mentioned in the introduction, the energy of the wavefunction can be extracted from a projected estimator (Eq. 1), which in this paper is simply obtained from a projection onto the Hartree–Fock determinant ($|\psi_T\rangle = |D_{\text{HF}}\rangle$). In addition, the value of E_S is varied throughout the simulation in order to maintain a constant, desired weight of walkers. At convergence, this value should fluctuate about the energy of the system, providing an alternative estimator for the energy based on the total growth rate of all the walkers in the system. More details on the specific implementation of these steps, and the derivation of this dynamic from the imaginary-time Schrödinger equation can be found in Refs. 10, 34, and 35. Furthermore, the systematically improvable *initiator* approximation is used exclusively throughout this work (sometimes denoted *i*-FCIQMC to distinguish it from the full method). This involves a dynamically truncated Hamiltonian operator, where spawning events to unoccupied determinants are constrained to be only allowed if they come from a determinant with a walker weight greater than n_{add} . This approximation can be systematically improved as the number of walkers increases, as increasing numbers of determinants fulfil

the criteria, and the sampled Hamiltonian therefore approaches the exact Hamiltonian. More details and benchmarking of this approximation can be found in Refs. 14 and 16.

A. Non-integer walker weights

Recent advances in the methodology were introduced by Petruzielo *et. al.*³⁵, where among other things, non-integer walker weights were introduced. This non-integer extension to FCIQMC requires a realization of Eq. 6 which allows walkers with non-integer weights to be spawned. This is done by applying the spawning and death processes continuously rather than discretely. By ‘continuous’ in this context, we mean that weights are assigned in a spawning/death process in a continuous framework, without the additional stochastic process to convert the resultant weight into an integer value for Δn_i . Applying both of these steps continuously removes the need for much of the random number generation in the code, and reduces the instantaneous fluctuations on any given determinant by allowing modification of its population by small fractions of a walker, rather than by whole walkers at a time. It should also be noted that when determining the number of *attempted* spawning events from a determinant with a non-integer number of walkers, the fractional part of the amplitude is used as a stochastic test for whether to perform an additional spawning step on top of $\lfloor n_i \rfloor$, to ensure the overall number of spawning attempts is still proportional to n_i , where $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ represent rounding down or up to the next integer value of a real number respectively.

However, with continuous spawning events automatically accepted (unless $|H_{ij}|$ is exactly zero), the number of successful spawning events is substantially increased. This requires more memory to store the increase in spawned walkers, but also slows the simulation down, as each spawning event must be put through some combination of communication and annihilation steps in order to transfer this information. It is therefore necessary to have some level of stochastic compression of the low-weighted walkers, to avoid storage costs for walkers which would quickly span much of the space.

In this work, the overhead of dealing with the additional low weighted walkers is controlled by two parameters. First, a subspace over which the non-integer walker weights are allowed is imposed. This is most simply done via a restriction of non-integer weights to only be allowed within a subspace defined by a given number of particle-hole excitations from the reference determinant (generally Hartree–Fock), determined by the cutoff parameter χ . This greatly minimizes the fluctuations in E_P , since any value of $\chi \geq 2$ will ensure that all determinants which contribute to $\langle E_P \rangle$ (with the Hartree–Fock trial wavefunction) are included in the non-integer space, which is finely resolved. Any spawning event out of this space requires an additional stochastic step to discretize

the walker amplitude to an integer quantity.

In addition, a minimum occupation threshold (N_{occ}) is defined, in order to ensure that low-weighted determinants do not proliferate, as was initially introduced in Ref. 35. After all annihilation events are complete, the instantaneous population of each determinant, n_i , is assessed in comparison to this threshold. If $n_i < N_{\text{occ}}$, its population is rediscrretised to either N_{occ} (with probability $\frac{n_i}{N_{\text{occ}}}$) or 0 (with probability $1 - \frac{n_i}{N_{\text{occ}}}$). Larger values of N_{occ} will therefore reduce the occupied space of determinants by depopulating more determinants with $n_i < N_{\text{occ}}$. Of course, values in excess of 1 correspond to a *coarser* representation of the wavefunction than the original integer algorithm in certain regions of the space, making such a choice inadvisable. Conversely, setting $N_{\text{occ}} = 0$ has no effect at all and returns us to the untenable scenario where a very large number of instantaneously occupied determinants with low weight must be simultaneously stored.

An alternative strategy is to restrict the total number of successful spawning events, which is known to be much higher in the non-integer algorithm. Many of these events may be propagating only small fractions of a walker at a time. For spawning events in particular, the cost of treating these low weight progeny in terms of storage and communication, can be substantial compared to their negligible impact on the progress of the simulation. The death events do not bear the same cost, as the populations are directly updated in the main list of determinants. Therefore, a minimum threshold value, κ , can be introduced for a continuous spawning event, with the advantage over N_{occ} that additional communication costs are avoided (these strategies can of course be used together). If p_{spawn} (as given in Eq. 7) is *greater* than κ , p_{spawn} walkers are spawned as a continuous event. If p_{spawn} is *less* than κ (i.e. a low weight event) κ walkers are spawned with probability $\frac{p_{\text{spawn}}}{\kappa}$, otherwise 0 walkers are spawned. By doing this, the disproportionate cost of treating progeny smaller than κ is saved.

It was found that the description of at least some parts of the space in this continuous fashion, allowing for finer resolution of these wavefunction amplitudes, was universally beneficial, with efficiency gains of up to 120 times over full use of integer walker weights. This was due to the reduction in stochastic noise, although the precise efficiency gain was significantly system-dependent. The change was also not found to have any negative effect on equilibration or serial correlation times. The gain in efficiency was found to be fairly consistent across values for χ , N_{occ} and κ , however, an approach with $\chi = 4$ (for a CISDTQ non-integer subspace) and a minimum occupation of 1 walker ($N_{\text{occ}} = 1$) was found to be advisable due to its more modest memory demands compared to a more complete non-integer walker space. This recommended approach typically resulted in 50% increased computational effort, and approximately twice as much memory than the integer method, which is more than offset by the gains in efficiency. These parameters will

be used throughout this work unless otherwise specified, with $\kappa = 0$. A more detailed analysis of this efficiency gain can be found in Ref. 36.

III. STOCHASTIC SAMPLING OF REDUCED DENSITY MATRICES

The second-order reduced density matrix (2-RDM) can be expressed in second quantization within a basis of spin orbitals as a rank-4 tensor, Γ ,

$$\Gamma_{pq,rs} = \langle \Psi | a_p^\dagger a_q^\dagger a_s a_r | \Psi \rangle \quad p > q, \quad r > s \quad (8)$$

where a_p^\dagger and a_p are creation and annihilation operators respectively³⁷. As the N -electron Hamiltonian is constructed only from 1- and 2-body operators, the 2-RDM representation contains sufficient information to exactly reproduce the total energy, as

$$E_{\text{RDM}} = \sum_{pq} h_{pq} \gamma_{pq} + \sum_{p>q,r>s} \Gamma_{pq,rs} \langle pq || rs \rangle + h_{\text{nuc}} \quad (9)$$

where the 1-RDM is represented as $\gamma_{pq} = \langle \Psi | a_p^\dagger a_q | \Psi \rangle$. In addition to accessing the total energy, the 2-RDM can also be used to compute other 1- and 2-body static properties of the system, provided that the appropriate operator can be projected into the space. This is achieved by computing the trace of the relevant 1- or 2-electron operator \hat{Q} with the RDM

$$\langle Q \rangle = \text{Tr}[\Gamma \hat{Q}] \quad (10)$$

where \hat{Q} is expressed in the same one-particle basis as Γ (Ref. 31). As such, the reduced density matrices are a powerful construct, allowing the calculation of a variety of *pure* expectation values with a highly compact representation of the relevant information in the N -electron wavefunction, without introducing any approximation into the total energy or other observable quantities^{29–31}. The reduced density matrices also form the basis of various measures of quantum entanglement³⁸, can be used to compute multireference explicit correlation corrections to reduce finite basis error^{32,39,40}, as well as provide the information for self-consistent optimisations of orbital spaces, as required for the complete active space self-consistent field (CASSCF) approach^{41,42}, and the density matrix embedding theory (DMET)^{43–45}.

For a FCI wavefunction expansion in a determinant basis with coefficients C_i , the 2-RDM can be written as

$$\Gamma_{pq,rs} = \sum_{ij} C_i C_j \langle D_i | a_p^\dagger a_q^\dagger a_s a_r | D_j \rangle \quad (11)$$

where the matrix element is non-zero only when D_j is connected to D_i by an $rs \rightarrow pq$ excitation, giving a value of ± 1 (repeated indices are also allowed). Diagonal terms in the RDMs can therefore be found as

$$\Gamma_{pq,pq} = \sum_{\{p,q\} \in i} (C_i)^2. \quad (12)$$

Therefore, within the FCIQMC dynamic, each determinant, $D_{\mathbf{i}}$, contributes to the 2-RDM via $C_{\mathbf{i}}^2$ summed into each of the $\frac{N(N-1)}{2}$ elements where orbitals p and q are occupied in $D_{\mathbf{i}}$, to generate the diagonal elements of the 2-RDM in a relatively simple fashion.

However, the explicit generation of all relevant determinant pairs in Eq. 11 for the off-diagonal contributions is prohibitively expensive ($\mathcal{O}[N^2M^2]$ per occupied determinant), and so we seek an algorithm which avoids this naive approach. This was considered in Ref. 32. As the 2-RDM involves contributions from pairs of determinants that are (at most) double excitations of one another, the existing spawning events that occur throughout the simulation to propagate the walker distribution in imaginary time already generate the required excitations, through the application of the Hamiltonian operator.

For each walker on determinant $D_{\mathbf{i}}$, the spawning step generates a target determinant, $D_{\mathbf{j}}$, which is always a single or double excitation of $D_{\mathbf{i}}$. If the spawning attempt is successful, the child walker is communicated to the parallel process that hosts $D_{\mathbf{j}}$, allowing its weight to be updated. The stochastic RDM method exploits this existing computational effort and communication to sample the RDM off-diagonal elements. To calculate the contribution to the unnormalized RDM from this determinant pair, the signed amplitude information of both determinants is required, which can be achieved by passing a small amount of additional information through the existing communication step with newly spawned walkers (an amplitude and the identity of the parent determinant, $D_{\mathbf{i}}$). After the communication step, the identity and signed weight of both determinants are then available on the same process. The identity of the orbitals that differ between them can then be easily calculated, and the contribution added to the relevant off-diagonal RDM element(s). Therefore, contributions to the off-diagonal RDM terms are only included when there is a successful spawning attempt between two determinants.

As the RDM contribution corresponding to $C_{\mathbf{i}}C_{\mathbf{j}}$ is only added in when there is a successful spawning event between the determinants, it must be rescaled to take into account the probability of this event occurring. The contribution added into these terms will therefore take the form

$$\frac{C_{\mathbf{i}}C_{\mathbf{j}}}{p_c(n_{\mathbf{i}})(\mathbf{j}|\mathbf{i})} \quad (13)$$

where $p_c(n_{\mathbf{i}})(\mathbf{j}|\mathbf{i})$ is the normalized probability of spawning at least one child (of any weight) onto $D_{\mathbf{j}}$ from $D_{\mathbf{i}}$ during the current iteration. This depends on factors such as the number of walkers on $D_{\mathbf{i}}$, and whether it is a stochastic or continuous spawning event. This can be calculated for non-integer weighted determinants as

$$p_c(n_{\mathbf{i}})(\mathbf{j}|\mathbf{i}) = 1 - ([n_{\mathbf{i}}] - n_{\mathbf{i}})\lambda^{[n_{\mathbf{i}}]} - (n_{\mathbf{i}} - [n_{\mathbf{i}}])\lambda^{[n_{\mathbf{i}}]} \quad (14)$$

and for determinants populated by integer numbers of walkers as

$$p_c(n_{\mathbf{i}})(\mathbf{j}|\mathbf{i}) = 1 - \lambda^{n_{\mathbf{i}}} \quad (15)$$

where $n_{\mathbf{i}}$ denotes the instantaneous walker weight on determinant $D_{\mathbf{i}}$ for the iteration in consideration, and λ is the probability of *not* spawning a walker (of any weight) from $D_{\mathbf{i}}$ to $D_{\mathbf{j}}$ in a single attempt. This value varies depending on what ‘type’ of spawning event is being attempted. Specifically,

$$\lambda_{\text{stoch.}} = 1 - \min(\tau|H_{\mathbf{j}\mathbf{i}}|, p_{\text{gen}}(\mathbf{j}|\mathbf{i})) \quad (16)$$

$$\lambda_{\text{cont.}} = 1 - p_{\text{gen}}(\mathbf{j}|\mathbf{i}) \quad (17)$$

$$\lambda_{\text{spawn cutoff}} = 1 - \frac{\tau|H_{\mathbf{j}\mathbf{i}}|}{\kappa} \quad \text{if } p_{\text{spawn}} < \kappa \quad (18)$$

where $\lambda_{\text{stoch.}}$ (Eq. 16) is to be used in the case of a stochastically realized integer spawning event, $\lambda_{\text{cont.}}$ (Eq. 17) is to be used for a continuous spawning event, and $\lambda_{\text{spawn cutoff}}$ (Eq. 18) is to be used if the continuous spawning truncation parameter, κ , is being used (see section II A) and $p_{\text{spawn}} < \kappa$, otherwise $\lambda_{\text{cont.}}$ should be used. It should be noted that the unbiasing factor $p_c(n_{\mathbf{i}})(\mathbf{j}|\mathbf{i})$ is not a constant, and is an implicit function of β since it depends on the instantaneous walker weight $n_{\mathbf{i}}$, for the iteration under consideration.

If multiple successful spawning events are registered between a specific $\{\mathbf{i}, \mathbf{j}\}$ pair in a single iteration, then it is necessary to only consider this to be a single contribution to the density matrix. In principle, if a pair of determinants contributes to the off-diagonal RDM element $\Gamma_{pq,rs}$, it will also contribute to the element $\Gamma_{rs,pq}$. However, in practice, we choose to only update one of these terms for a spawning event, with the other sampled by events $D_{\mathbf{j}} \rightarrow D_{\mathbf{i}}$. The difference between the two terms, related by the required Hermiticity property of the sampled matrix, can then be used as a metric of the quality of off-diagonal sampling of the RDM, before it is averaged to become strictly Hermitian at the end of the calculation (numerically equivalent to including both contributions at the accumulation stage).

These stochastic reduced density matrices do not satisfy the trace relation, given (for the 2-RDM) by

$$\sum_{pq} \Gamma_{pq,pq} = \frac{N(N-1)}{2} \quad p > q, \quad (19)$$

where N is the number of active electrons. This is because the $n_{\mathbf{i}}$ values are not normalized determinant coefficients. This is more difficult to account for directly, as normalising the full wavefunction requires average sign values for all determinants, not just those instantaneously occupied. Instead, the 2-RDM is normalised directly at the end of the simulation to enforce the trace relationship in Eq. (19), which we are free to do as this normalization is a freedom in the sampling of the wavefunction. Further N-representability constraints would be hard to apply, since these generally correspond either to structural relationships between higher-ranked RDMs (which we do not have access to in this scheme) and lower ranked ones, or constraints on the spectrum of the reduced density matrices which would be difficult to

constrain at the sampling stage^{46–48}. Furthermore, we believe in this scheme that the RDMs obtained should as faithfully as possible reflect the wavefunction which is sampled, rather than imposing additional constraints, which would result in derived properties not directly corresponding to expectation values of the sampled wavefunction. It should be stressed that if the walker distribution is sampling the correct wavefunction, then all N -representability conditions will be exactly satisfied upon appropriate averaging in imaginary time.

The technique outlined above allows for an accumulation of an approximate RDM, alongside the stochastic dynamic of the FCIQMC wavefunction. However, there are a couple of subtleties which need to be considered. The first concerns the fact that off-diagonal contributions to the RDM arising from determinant pair $\{D_i, D_j\}$ are only included in the event of a successful spawning event from one to the other. The probability of this occurring is proportional to the Hamiltonian matrix element connecting the two determinants (in the case of a stochastic spawning event), which means that determinant pairs whose spawning events are discretised and connected via large Hamiltonian matrix elements will be sampled more frequently than weakly coupled determinants. This can result in sampling difficulties if two significantly weighted determinants are *not* in the non-integer walker space, and are connected via very small matrix elements, since a contribution to the RDM may be poorly sampled, or even entirely omitted. It should be stressed that continuous spawning connections are not sampled proportionally to the Hamiltonian matrix element, and so are not affected.

A specific example where issues can arise concerns the connections to single excitations of the Hartree–Fock determinant (D_{HF}) in a canonical basis, whose connecting matrix elements are therefore numerically zero by Brillouin’s theorem. To account for this, connections between the reference (generally Hartree–Fock) determinant and its excitations (both single and double excitations) are taken into account explicitly. This is simple to do since the weight on the reference determinant is known to all parallel processes, as it is used for the calculation of the projected energy estimate, E_P . Since the contribution from these pairs of determinants is performed exactly (which are likely to be the main contribution, especially in relatively weakly correlated systems), their contributions are not dependent on a successful spawning event between the two.

However, it is still possible that other very small Hamiltonian matrix elements connect substantially weighted determinants, and a sampling error could then arise. This was investigated by including additional spawning events between determinants. These events would spawn walkers between determinants proportionally to a function of the *inverse* of the Hamiltonian matrix element between them, or a specific lower-bound cutoff. These spawned walkers would not adjust the walker amplitude on determinants they were spawned to, or affect the simulation trajectory, but rather would simply confer

information about the respective determinant parent and its amplitude (with appropriately adjusted p_c) in order to include the contribution to the off-diagonal elements of the RDMs when the Hamiltonian matrix element between the two determinants was small. This was found to provide negligible improvement across a range of systems as this sampling error was small, and therefore we consider this unnecessary and do not pursue it further here.

Despite this, the quality of the sampled density matrices remains poor, and convergence to the exact density matrix and derived properties with increasing walker number is slower than for properties such as the projected energy of the system (see results in Ref. 32). This is due to a remaining significant error beyond simply a manifestation of insufficient sampling time or initiator error of the wavefunction. This remaining bias can be summarized by the observation that the appropriate contribution to the density matrix elements, $\langle n_i(\beta) \rangle_\beta \langle n_j(\beta) \rangle_\beta$, is in fact approximated by $\langle n_i(\beta) n_j(\beta) \rangle_\beta$. This ignores the non-zero (co)variance between the two amplitudes, which (especially for diagonal contributions) can be significant, and results in a biased sampling, even if the averaged walker amplitudes themselves are unbiased. The root cause is therefore the fact that the two amplitudes which contribute to the density matrix element are correlated, with the dominant error unsurprisingly coming from the diagonal elements where the instantaneous amplitudes are perfectly correlated, and where the error is of a single sign, removing the possibility of error cancellation in the neglected (co)variance.

Computing appropriately averaged values of the determinant coefficients over the simulation, $\langle n_i(\beta) \rangle_\beta$, which would remove the instantaneous correlation between the walker populations, has prohibitive memory demands as it involves storage of determinants that are occupied at *any* stage of the simulation, even if their instantaneous population is zero. A simple approach to reduce the bias is to average the walker populations during the period over which the determinant is occupied and use this as an approximate value for the $\langle n_i(\beta) \rangle_\beta$ determinant amplitude for the density matrix contribution (as was performed for Ref.³² and the ‘biased’ comparison results in section IV). However, since this contribution is not averaged over periods while the determinant is unoccupied, the density matrix is still a biased expectation value of the sampled wavefunction. Numerical results suggest that although this approach will tend to the exact density matrix and associated properties in the large walker limit, the convergence to the exact result requires far more walkers compared to the projected energy estimate.

In a recent article⁴⁹ Blunt *et al.* introduced the ‘density matrix quantum Monte Carlo’ (DMQMC) method. In this, the *full* N -electron density matrix is sampled stochastically. From this quantity, lower-rank reduced density matrices can be obtained via ‘tracing out’ of additional degrees of freedom. However, this approach contrasts with the approach taken here, where only the

wavefunction is sampled, and we seek to directly find the reduced density matrix which corresponds to this sampled wavefunction. However, we draw inspiration from the DMQMC approach, since they considered the use of a replica trick to overcome a similar difficulty relating to the evaluation of a quadratic function of a stochastically sampled distribution (this time of the density matrix rather than wavefunction) – the Renyi S_2 entanglement entropy. In this, two uncorrelated walker samples were propagated independently, thus rigorously removing any covariance between the sampled distributions. They further suggest that this approach might be useful for sampling expectation values, such as in Eq. 4, directly within FCIQMC. This approach was subsequently used in Ref. 50, in order to compute *dynamical* expectation values, including the Green’s functions and arbitrary spectra. In the rest of this article we take up this idea to calculate RDMs within FCIQMC and demonstrate that an unbiased sampling can indeed be achieved, dramatically improving the convergence over the biased approach.

IV. UNBIASED FCIQMC DENSITY MATRIX REPLICA SAMPLING

An unbiased RDM sampling can be achieved via two independent walker populations (replicas), whose instantaneous weights on a determinant D_i will be denoted $n_i^{(1)}$ and $n_i^{(2)}$ for the first and second walker replicas respectively. In this implementation, the replicas are propagated with independent random number strings, and each determinant with a walker weight in either of the two replicas is stored in the data structures outlined in Ref. 34, in the same fashion as handling complex walker amplitudes in Ref. 19. This allows for the weight on a given determinant over both replicas to be extracted without any additional computational overhead.

Spawning, death and annihilation events are then applied to each population independently and two sets of statistics gathered, including measures of E_P and E_S , with $E_S^{(1)}$ and $E_S^{(2)}$ used in the respective death probabilities for the replicas. The two simulations are therefore completely uncorrelated, apart from sharing the same initial conditions, from which they quickly decorrelate after equilibration. Although this effectively doubles the computational resources required per iteration, there are two independent estimates of all quantities of interest, and so by using this additional information, the random errors in these quantities are reduced by a factor of $\sqrt{2}$. This results in the calculation requiring only half the time. Computational overheads therefore only effectively result from increased memory requirements from the additional replica population.

Uncorrelated and unbiased contributions to elements of the reduced density matrices can then be found by ensuring that any product of determinant populations is

computed from the amplitudes over both replicas. Contributions to the diagonal elements of the RDMs can be gathered each iteration from all instantaneously occupied determinants, by accumulation of $n_i^{(1)}(\beta) n_i^{(2)}(\beta)$ to the $\frac{N(N-1)}{2}$ relevant diagonal elements. Off-diagonal elements are accumulated through successful spawning events, as detailed in section III, ensuring that the two determinant weights n_i and n_j come from different populations to ensure an unbiased sampling. However, there are two possibilities with this.

From consideration of a determinant transition from replica 1, $D_i \rightarrow D_j$, which occurs with probability $p_c(\mathbf{j}|\mathbf{i})^{(1)}$, two possible contributions could be considered;

$$\text{Forwards} \quad \frac{n_i^{(1)} n_j^{(2)}}{p_c(\mathbf{j}|\mathbf{i})^{(1)}} \quad (20)$$

$$\text{Backwards} \quad \frac{n_i^{(2)} n_j^{(1)}}{p_c(\mathbf{j}|\mathbf{i})^{(1)}}. \quad (21)$$

It is important that only the ‘forwards’ contribution is considered for the spawning events from replica 1 (Eq. 20), since including the latter ‘backwards’ contribution (Eq. 21) introduces a bias into the sampling. This can be rationalized from consideration of an iteration where $n_i^{(1)} = 0$ and $n_i^{(2)} \neq 0$. In this case, there would be non-zero $n_i^{(2)} n_j^{(1)}$ contributions which would never be sampled, because there are no walkers on site i in the first replica to spawn from. Instead the ‘backwards’ contribution should be sampled only from transitions within replica 2, with the denominator changed to $p_c(\mathbf{j}|\mathbf{i})^{(2)}$, and appropriately averaged with the forwards contribution. However, for the results presented in this paper, the contributions from only one replica were considered, indicating that a further reduction of the error bars in the density matrix elements by a factor of $\sqrt{2}$ is trivially possible without additional cost.

Finally, as described in section III, the off-diagonal contributions that involve direct connections to the reference determinant (generally D_{HF}) can be included explicitly, with the knowledge of n_{HF} for both replicas known to all processes each iteration. This will ensure the best quality sampling for these generally important contributions, and avoid any sampling issues arising from Brillouins theorem. Again, the sampled density matrices are normalized to fulfil the appropriate trace relations, as well as made Hermitian at the end of the simulation (which is only exactly achieved without this in the limit of long sampling time), which ensures that $\Gamma_{pq,rs} = \Gamma_{rs,pq}^*$ for the 2-RDM.

A simple comparison between the original, biased approach, and the replica sampling of the density matrices in Fig. 1 shows the striking improvement in the quality of the properties. It can be seen that the convergence of the quasi-variational energy expectation value, E_{RDM} , is now comparable to that of the non-variational estimate, E_P (or faster in some instances). The E_{RDM} estimate is denoted ‘quasi-variational’, since it is only rigorously

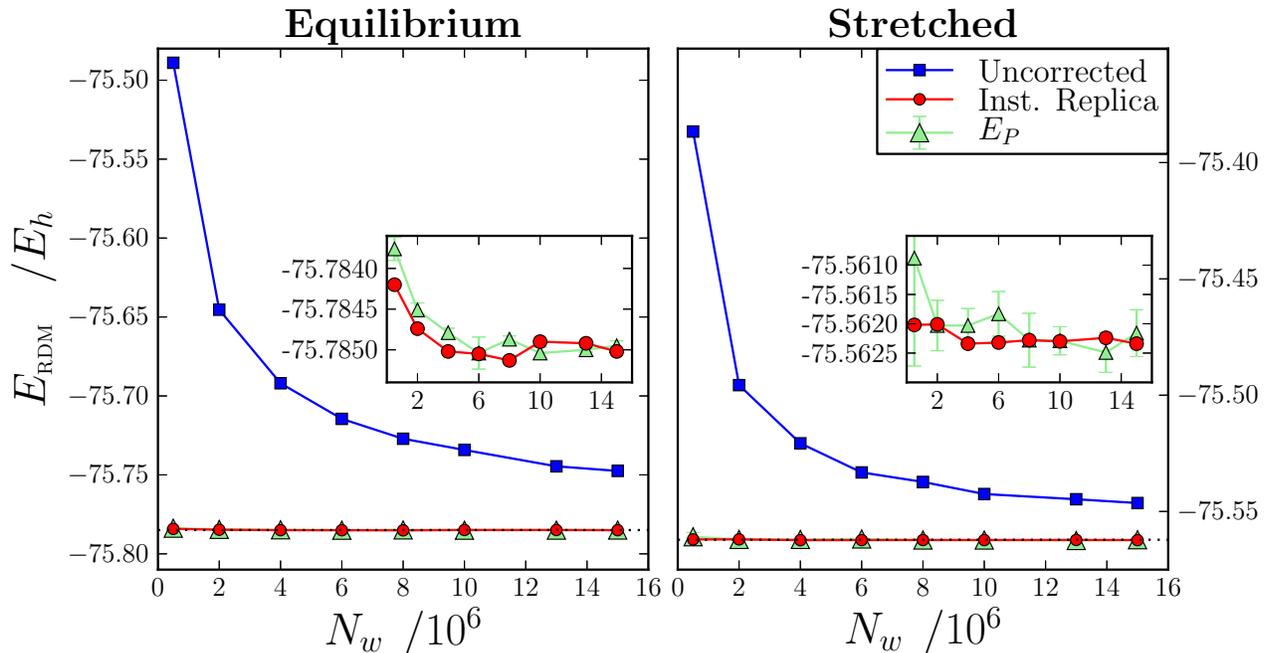


FIG. 1. Comparison of the convergence with number of walkers (N_w) for the energy estimate calculated from the sampled density matrices (E_{RDM} as given in Eq. 9), for C_2 in a cc-pVTZ basis. The Hilbert space size is $\mathcal{O}[10^{10}]$ many-body functions. ‘Uncorrected’ refers to the biased sampling detailed in section III, while the expectation values from the unbiased replica-sampled RDMs of section IV are denoted ‘Inst. Replica’. Also for comparison is the projected energy estimate, E_P . This system is identical to the one studied in Ref 32, which used the biased sampling algorithm, although all values have been newly generated here. Uncorrected RDMs were calculated with the integer i-FCIQMC algorithm (consistent with Ref. 32), whilst replica-sampled RDMs use the non-integer walker weights as detailed in section IIA, with $\chi = 4$, $N_{\text{occ}} = 1$, to represent the best quality RDM available with the techniques presented in this paper. Errorbars are not included for E_{RDM} estimates, since only one calculation was performed. The use of non-integer walker weights does not generally remove the systematic error, and so the improvement in E_{RDM} between the two sampling techniques can be considered a result of the use of the unbiased replica sampling. Remaining systematic error at low walker numbers can be attributed to the convergence of the initiator approximation. This manifests more strongly in the equilibrium results than the multiconfigurational stretched case.

variational in the long sampling limit. For short sampling times, it is possible that the replicas have sampled different wavefunctions within their stochastic errors, and therefore the N -representability conditions of the RDMs will not be strictly satisfied. Non-variational energies are therefore possible within the random errors of the sampling.

In the limit of sufficient sampling of imaginary time, but where the number of walkers is not sufficient to entirely converge the initiator error of the FCIQMC method, E_{RDM} will be variational, but is not in general the same as E_P . These two energy estimates will only rigorously be the same where the sampled wavefunction represents an eigenstate. Therefore, this quasi-variational estimate of the energy may have a different convergence profile with N_w compared to the standard energy estimates. This might provide a clearer measure of decay of the initiator error, as well as potentially being more amenable to variational extrapolation of this energy to the infinite walker limit to remove the effects of the initiator error. This will be investigated in future work.

The convergence to this large walker limit of the E_{RDM} expression can be assessed via comparison to FCI results, as shown in Fig. 2, where a smaller system than that of Fig. 1 allows the exact 2-RDM to be calculated. This indicates that the E_{RDM} expression appears to provide a more rapid convergence to the large walker limit in the case of more multiconfigurational wavefunctions, simulated by the stretching of the triply-bonded nitrogen molecule to $4.2 a_0$. This is unsurprising, since the projected energy expression used a trial wavefunction which was simply the Hartree–Fock determinant, whose weight in the FCI expansion diminishes rapidly as the bond is stretched⁵¹. The use of a multiconfigurational trial wavefunction would improve the projected energy estimate in these cases. However, it is clear that in these cases, E_{RDM} provides a good estimate of the energy, with a far larger number of spawning events in the space directly contributing to the energy (all to a determinant with a non-zero population in the other replica), rather than just the distribution of a small fraction of walkers on the reference determinant and its direct excitations.

This improvement of the E_{RDM} energy estimator over

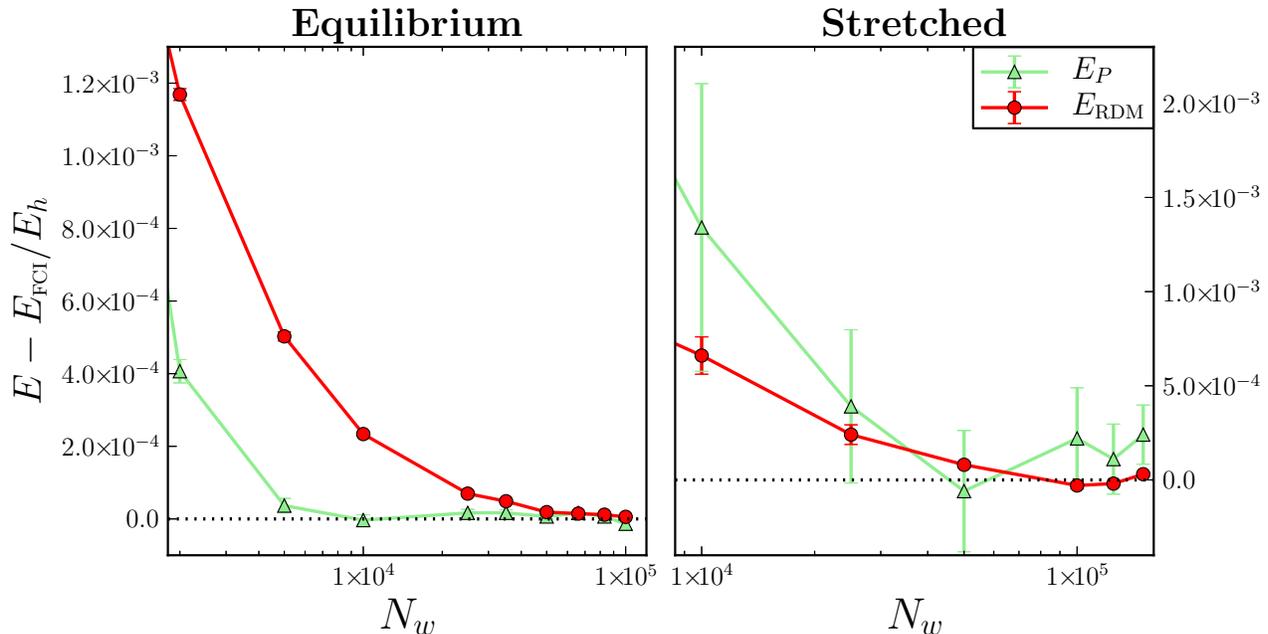


FIG. 2. Convergence of E_{RDM} and E_P with N_w for equilibrium (1.094Å) and stretched (4.2a₀) N₂ in a cc-pVDZ basis. E_{RDM} was calculated with the instantaneous replica sampling method, using non-integer coefficients. $\tau = 10^{-3}$ a.u.⁻¹ and 8 electrons were frozen. The Hilbert space size is $\mathcal{O}[10^5]$ many-body functions. Convergence rate is seen to be substantially improved (comparable to or faster than the projected energy estimate) for more multiconfigurational systems.

the E_P value is again borne out for multiconfigurational problems when considering not the number of walkers required, but the sampling time required for the random errors associated with the energy estimate to fall to an acceptable value. This can be observed in Fig. 3, where the convergence of the energy for a given number of iterations is in general faster than the corresponding projected energy estimate. The convergence of other properties derived from the sampled density matrices will be investigated in section V, after a consideration of the computational costs associated with sampling of the density matrices.

A. Computational Efficiency Considerations

The replica-sampled RDM method achieves an unbiased sampling of the 2-RDM by adding in instantaneous contributions $n_i^{(1)/(2)} \times n_j^{(2)/(1)}$ throughout the simulation. Although there is an associated computational cost for running another independent replica of the walker population, this cost is largely offset by the improvement in the statistics that result compared to running twice as long. Therefore, the dominant additional overhead for propagation of the additional replica comes from the associated memory requirements. The accumulation of the 2-RDM also has associated costs, since the memory required is of size $\mathcal{O}[M^4]$. Since this is of the same size as the storage of the two-electron integrals required for the

calculation and also stored in memory, it is currently not a bottleneck. However, eventually these memory costs may be required to be distributed among the computational nodes³⁴.

The dominant computational cost in sampling the density matrix comes from the accumulation of the diagonal matrix elements, as each occupied determinant requires an $\mathcal{O}[N^2]$ operation to identify and update all the necessary 2-RDM elements each iteration. Therefore, any approach that reduces the frequency at which these contributions need to be added in will reduce the computational cost associated with accumulating the RDM, with increasing returns at large N . This can be achieved with an approach whereby we store and calculate a running average of a determinant's weight within a defined block of iterations, and only add in its diagonal RDM contributions (and connections to reference determinant, if applicable) once at the end of these blocks, weighted by the number of iterations in that block. In so doing, contributions to the diagonal elements and D_{HF} connections can be summed in much less frequently without any loss of information.

However, the implementation of such an averaging technique requires careful thought to ensure that no systematic bias is introduced into the RDM contributions, whilst making the best use of the available data in the simulation. This dictates that 'blocks' of iterations over which the weight on a determinant is averaged should start whenever a determinant has a change in occupation status (i.e. becomes newly unoccupied or newly oc-

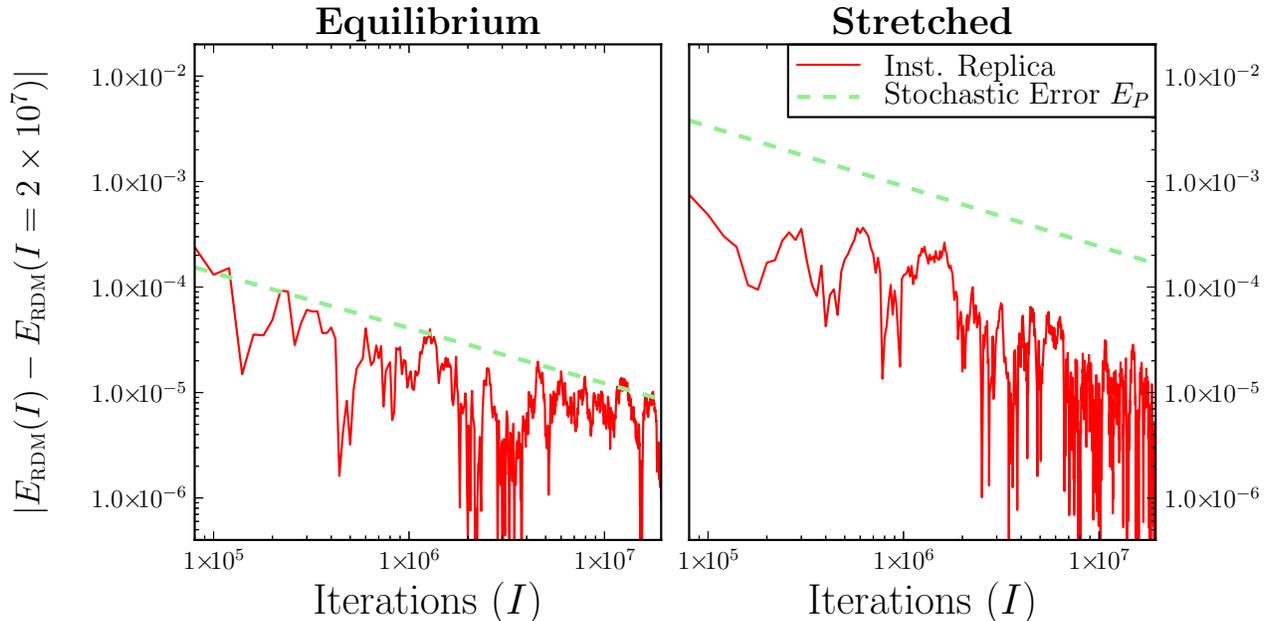


FIG. 3. Convergence of errors in E_{RDM} and E_P with Monte Carlo iterations for equilibrium and stretched N_2 . E_{RDM} values are calculated with the replica sampling method and non-integer coefficients. The values at each accumulation time are then plotted relative to the value at $I = 2 \times 10^7$, to give an estimate of the random sampling error in E_{RDM} with iterations I . The dashed green line represents the random error in E_P calculated over the same number of iterations and simulation setup. This was calculated by conducting a blocking analysis on data sets of differing length, and fitting to these data points⁵². The plotted fit lines for the E_P estimate have the form $\mathcal{O}[I^{-0.52}]$ and $\mathcal{O}[I^{-0.57}]$, consistent with the expected $\mathcal{O}[I^{-0.5}]$ scaling of stochastic errors in Monte Carlo methods. It can be seen that the error in the E_{RDM} estimate from the accumulated RDMs is (at all iteration values) less than the corresponding E_P energy, while following the same $\mathcal{O}[I^{-0.5}]$ rate of convergence. This indicates that the random errors in the E_{RDM} energy are equivalent, or smaller, than the corresponding E_P value for a given number of iterations, with the stretched case showing a greater improvement over E_P .

cupied) *in either replica population*. This ensures that the only blocks with a non-zero contribution to the RDM exactly correspond with the set of iterations that would have produced a non-zero contribution when adding in snapshots instantaneously every iteration.

This averaging approach retains the unbiased nature of the RDM sampling, but has a significant advantage in terms of simulation time, as contributions to the diagonal elements and D_{HF} connections are only required at block boundaries (and for the diagonal contribution, only when both replicas have been occupied in the block), rather than every iteration. Therefore, for a given determinant, its contributions to the RDM diagonal elements, which are added in only at the end of each averaging block, now take the form

$$\langle n_{\mathbf{i}}^{(1)} \rangle_a \langle n_{\mathbf{i}}^{(2)} \rangle_a I(D_{\mathbf{i}})_a \quad (22)$$

where $\langle n_{\mathbf{i}}^{(1)} \rangle_a$ is the average weight on $D_{\mathbf{i}}$ in population 1 within averaging block a , and $I(D_{\mathbf{i}})_a$ is the number of iterations within that averaging block. This removes what rapidly becomes the dominant cost as the number of electrons in the system increases. Similarly, if the given determinant is a single or double excitation of D_{HF} , the off-diagonal contributions from this connection are added

in at the same time as the diagonal contributions, taking the form:

$$\langle n_{\text{HF}}^{(1)} \rangle \langle n_{\mathbf{i}}^{(2)} \rangle_a I(D_{\mathbf{i}})_a \quad (23)$$

where $\langle n_{\text{HF}}^{(1)} \rangle$ is the measure of the average population on D_{HF} up to the iteration where the contribution is included.

Although this averaging approach can also be shown to achieve an unbiased sampling of the diagonal elements and the explicitly included D_{HF} connections, the presence of serial correlation in determinant amplitudes creates a problem for the remaining stochastically sampled off-diagonal contributions if these averaged coefficients are used. Therefore, off-diagonal elements must continue to be accumulated using instantaneous determinant weights as detailed previously. However, this is not a particular burden since for the off-diagonal elements there is only one contribution (for sampled double excitations), or $\mathcal{O}[N]$ (if a single excitation is sampled).

V. MOLECULAR PROPERTIES

It is important not to simply consider the density matrix energy estimator when benchmarking the quality of the RDMs, since the action of contracting the density matrices with the Hamiltonian will place emphasis on parts of the RDM matrix, while being relatively insensitive to others. Indeed, the algorithm dictates that it preferentially samples those elements whose contribution to the variational energy estimator is largest. In addition, the primary motivation for the sampling of the density matrices is for the computation of unbiased molecular properties derived from the density matrices, whose value could not be computed in an unbiased fashion from a projected estimator.

Figures 4 and 5 show various properties derived from the one- and two-body sampled density matrices, compared to exact results from density matrices computed from a FCI diagonalization of the system. Both an equilibrium (relatively single reference), and stretched triple bond (strongly multiconfigurational) are shown, to highlight any potential differences arising from the sampling of the RDMs within different wavefunction characteristics. Also, both the instantaneous replica RDM sampling, as well as the cheaper but still unbiased sampling of averaged determinantal weights for diagonal RDM elements are shown. These are all within random errorbars of each other, confirming the equivalence of these two approaches, and the fact that the averaging of determinant weights should always be performed.

The properties sampled in these results include the spin quantum number of the sampled wavefunction⁴⁸, and two differing corrections to the wavefunction ameliorating the basis set incompleteness in the representation, proposed by Valeev *et al.*^{40,53,54}. The first ($[2]_{R12}$) approximately corrects for the incompleteness in the two-electron form of the correlation hole around each coalescence point⁴⁰, while $[2]_S$ corrects for the incompleteness relating to the single-particle orbital representation⁵³. Although the $[2]_S$ correction was originally proposed in the context of accelerating convergence with respect to the external space within CASSCF calculations, in this context it is used without orbital optimization, with the entire orbital space considered as occupied. This can cause convergence difficulties and intruder states as $H^{(0)}$ does not necessarily separate between the orbital and CABS spaces. However, whether this quantity is a robust correction for single-particle basis incompleteness without combining it with orbital optimization is separate to the use of it as a metric for analysis of the convergence of the parent density matrix, for which it is primarily used in this study.

These quantities, as well as other explicit correlation corrections, have been applied to FCIQMC wavefunctions previously^{32,55}. It was found that these two properties were far more sensitive than E_{RDM} on the quality of the sampled off-diagonal 2-RDM elements, and so provided a complementary test of the sampling quality.

Nevertheless, the errors in these quantities is an order of magnitude smaller than E_{RDM} , and converges at least as rapidly to the exact result. The error in these quantities is expected to solely result from initiator error in the sampled wavefunction, which decreases as observed as the number of walkers increases, as all other sources of systematic error from the density matrices have been removed^{56,57}. The random error in the calculations is also very small, having been estimated from independent calculations.

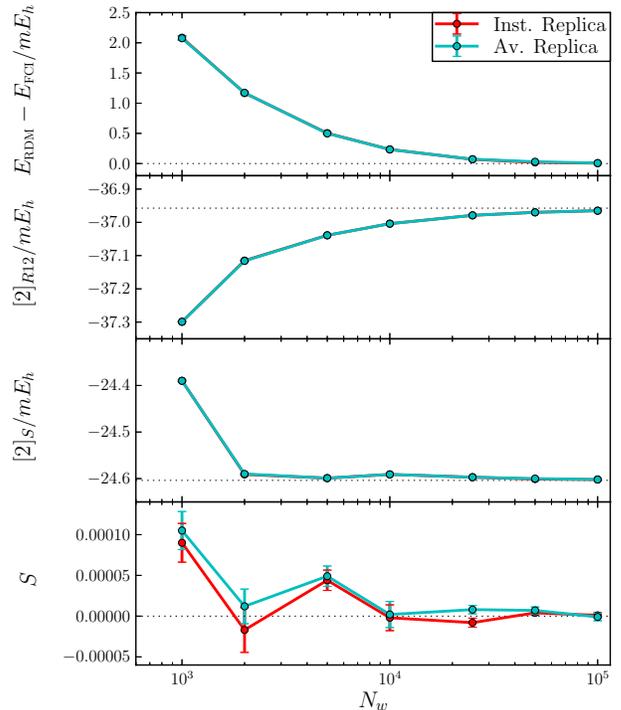


FIG. 4. Comparing instantaneous and averaged diagonal RDM replica sampling methods for convergence of selected properties with number of walkers, N_w , for equilibrium (1.094Å) N_2 cc-pVDZ. Both approaches are unbiased, and are shown to be equivalent within their random errorbars (most are entirely overlaid). Plotted is E_{RDM} , as well as $[2]_{R12}$ and $[2]_S$ explicit correlation energy corrections (see Refs. 40 and 54) and the value of the spin quantum number, S , of the sampled wavefunction⁴⁸. These errorbars are calculated from five independent calculations, but are often too small to be seen. Dotted lines give the exact FCI value for each quantity, while stochastic RDMs were accumulated over $\sim 5 \times 10^6$ iterations. 8 electrons are frozen, with $\tau = 10^{-3}$ a.u.⁻¹

VI. SUMMARY AND CONCLUSIONS

In summary, we propose the replica-sampled RDM method for computing one- and two-body reduced density matrices from a stochastic FCIQMC sampling of the wavefunction. The replica sampling allows for the elimination of systematic errors due to the quadratic de-

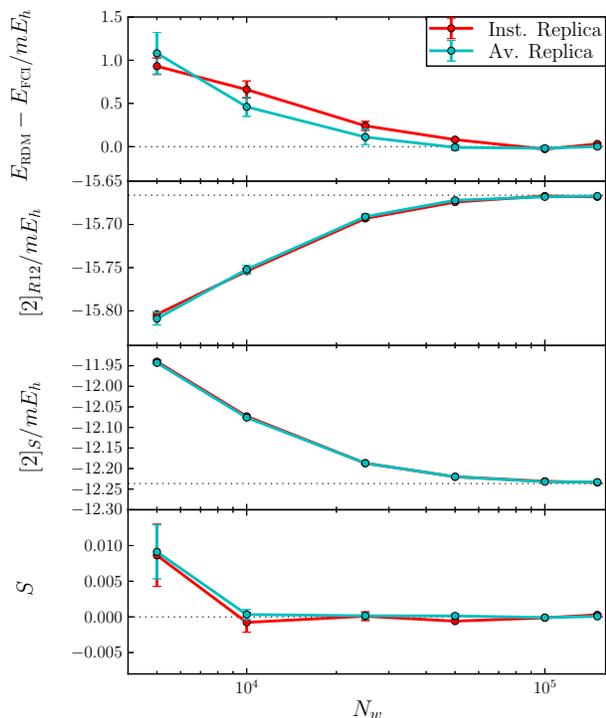


FIG. 5. This plot details equivalent results to Fig. 4, but for the more multiconfigurational, stretched N_2 molecule ($4.2a_0$). Note the reduction in the error in E_{RDM} as the bond is stretched. This contrasts in general with E_P , which degrades as the wavefunctions become less compact and more difficult to sample in these multiconfigurational cases⁵⁸.

pendence of the density matrix elements on the sampled wavefunction, from which a previous implementation suffered badly. The resultant reduced density matrices have no systematic error besides the initiator error present in the inherent sampling of the wavefunction itself, which can be systematically improved by increasing overall walker number. The accumulation of the RDMs runs alongside the sampling of the wavefunction, making use of the two-body operator moves of FCIQMC, therefore adding only modest additional computational overheads for the calculation of the density matrices.

Averaged determinant weights are used for the contributions to the RDM diagonal elements and D_{HF} connections, therefore avoiding a potentially costly $\mathcal{O}[N^2]$ step for each occupied determinant each iteration, without additional errors being incurred. However, instantaneous weights are used for the stochastic sampling of the remaining off-diagonal contributions, retaining the unbiased qualities of the original all-instantaneous replica sampling approach. The properties derived from these density matrices are shown to converge with increasing walker number to exact results at a comparable rate to the standard projected energy estimators, as the initiator error is reduced.

This is also true for the quasi-variational energy estimator derived from the density matrices (strictly varia-

tional in the large imaginary-time limit). This estimate is found to be an accurate alternative to the projected estimate, especially powerful in strongly correlated, multiconfigurational wavefunctions, where the absence of a large weighted wavefunction space to project onto can result in large random errors in this approach. Furthermore, the quasi-variational nature of the value may lend itself to extrapolation techniques for the systematic initiator error.

On the face of it, it would appear that the scaling of the replica-sampled reduced density matrix will be poor with respect to increasing system size, as it requires spawning events onto a determinant occupied in the other replica in order to accumulate off-diagonal contributions. However, it is precisely this type of event which is required for the annihilation part of the algorithm, crucial for the suppression of the sign-problem in the space (aided by the initiator conditions). Because of this, we do not anticipate the scaling to be particularly worse than that of the fundamental ability of the algorithm to resolve the correct wavefunction. This precise scaling is expected to be dependent on the expectation values being probed, as well as the system under investigation, and further studies are underway to observe the accuracy of the properties as the system sizes increase.

Finally, it should be noted that it is straightforward to combine this approach with more recent developments and extensions of FCIQMC, such as the semi-stochastic sampling of the space³⁵, extraction of excited states^{59,60}, and additional symmetries such as time-reversal symmetry and (angular) momentum symmetries which have been constrained in the walker dynamics¹⁴. We now plan on turning our attention to larger systems, and an assessment of the accuracy of other properties, including nuclear forces and electrical moments of molecules.

VII. ACKNOWLEDGEMENTS

GHB would like to gratefully acknowledge financial support from the Royal Society via a University Research Fellowship. This work has been supported by EPSRC grant number EP/J003867/1.

- ¹S. M. Rothstein, *Can. J. Chem.* **91**, 505 (2013).
- ²A. East, S. Rothstein, and J. Vrbik, *J. Chem. Phys.* **89**, 4880 (1998).
- ³P. Reynolds, *J. Chem. Phys.* **92**, 2118 (1990).
- ⁴R. Barnett, P. Reynolds, and W. Lester, *J. Comput. Phys.* **96**, 258 (1991).
- ⁵R. Barnett, P. Reynolds, and W. Lester, *J. Chem. Phys.* **96**, 2141 (1992).
- ⁶P. Langfelder, S. Rothstein, and J. Vrbik, *J. Chem. Phys.* **107**, 8525 (1997).
- ⁷S. Baroni and S. Moroni, *Phys. Rev. Lett.* **82**, 4745 (1999).
- ⁸R. Gaudoin and J. M. Pitarke, *Phys. Rev. Lett.* **99**, 126406 (2007).
- ⁹R. Gaudoin and J. M. Pitarke, *Phys. Rev. B.* **81**, 245116 (2010).
- ¹⁰G. H. Booth, A. J. W. Thom, and A. Alavi, *J. Chem. Phys.* **131**, 054106 (2009).
- ¹¹G. H. Booth and A. Alavi, *J. Chem. Phys.* **132**, 174104 (2010).

- ¹²J. S. Spencer, N. S. Blunt, and W. M. C. Foulkes, *J. Chem. Phys.* **136**, 054110 (2012).
- ¹³D. Cleland, G. H. Booth, and A. Alavi, *J. Chem. Phys.* **132**, 041103 (2010).
- ¹⁴G. H. Booth, D. Cleland, A. J. W. Thom, and A. Alavi, *J. Chem. Phys.* **135**, 084104 (2011).
- ¹⁵D. Cleland, G. H. Booth, C. Overy, and A. Alavi, *J. Chem. Theory Comput.* **8**, 4138 (2012).
- ¹⁶D. M. Cleland, G. H. Booth, and A. Alavi, *J. Chem. Phys.* **134**, 024112 (2011).
- ¹⁷J. J. Shepherd, G. E. Scuseria, and J. S. Spencer, *Phys. Rev. B* **90**, 155130 (2014).
- ¹⁸C. Daday, S. Smart, G. Booth, A. Alavi, and C. Filippi, *J. Chem. Theory Comput.* **8**, 4441 (2012).
- ¹⁹G. H. Booth, A. Grueneis, G. Kresse, and A. Alavi, *Nature* **493**, 365 (2013).
- ²⁰J. J. Shepherd, G. Booth, A. Grüneis, and A. Alavi, *Phys. Rev. B* **85**, 081103 (2012).
- ²¹J. J. Shepherd, G. H. Booth, and A. Alavi, *J. Chem. Phys.* **136**, 244101 (2012).
- ²²M. Kolodrubetz and B. K. Clark, *Phys. Rev. B* **86**, 075109 (2012).
- ²³P. Pulay, *WIREs Comput. Mol. Sci.* **4**, 169 (2014).
- ²⁴R. P. Feynman, *Phys. Rev.* **56**, 340 (1939).
- ²⁵W. M. C. Foulkes, L. Mitras, R. J. Needs, and G. Rajagopal, *Rev. Mod. Phys.* **73**, 33 (2001).
- ²⁶We note that there has been recent work to also sample from a mixed distribution within an FCIQMC framework²². However, as mentioned, this difference is not fundamental to the inherent bias or approach presented in this paper, which can also be employed within mixed distribution sampling of FCIQMC.
- ²⁷J. S. Spencer, N. S. Blunt, and W. M. Foulkes, *J. Chem. Phys.* **136**, 054110 (2012).
- ²⁸R. M. Lee, G. J. Conduit, N. Nemeč, P. López Ríos, and N. D. Drummond, *Phys. Rev. E* **83**, 066706 (2011).
- ²⁹J. Coleman, in *Reduced-Density-Matrix Mechanics*, Advances in Chem. Phys., Vol. 134, edited by D. Mazziotti (John Wiley & Sons Inc., New Jersey, 2007).
- ³⁰D. Mazziotti, in *Reduced-Density-Matrix Mechanics*, Advances in Chem. Phys., Vol. 134, edited by D. Mazziotti (John Wiley & Sons Inc., New Jersey, 2007).
- ³¹K. Blum, *Density Matrix Theory and Applications* (Springer-Verlag, Berlin, 2012).
- ³²G. H. Booth, D. Cleland, A. Alavi, and D. P. Tew, *J. Chem. Phys.* **137**, 164112 (2012).
- ³³M. H. Kolodrubetz, J. S. Spencer, B. K. Clark, and W. M. Foulkes, *J. Chem. Phys.* **138**, 024110 (2013).
- ³⁴G. H. Booth, S. D. Smart, and A. Alavi, *Mol. Phys.* **112**, 1855 (2014).
- ³⁵F. R. Petruzielo, A. A. Holmes, H. J. Changlani, M. P. Nightingale, and C. J. Umrigar, *Phys. Rev. Lett.* **109**, 230201 (2012).
- ³⁶C. Overy, *Reduced Density Matrices and Stochastic Quantum Chemistry*, Ph.D. thesis, University of Cambridge (2014).
- ³⁷T. Helgaker, P. Jørgensen, and J. Olsen, *Molecular Electronic-Structure Theory* (Wiley, 2000).
- ³⁸V. Popkov and M. Salerno, *Int. J. Mod. Phys. B* **26**, 1243009 (2012).
- ³⁹L. Kong, F. A. Bischoff, and E. F. Valeev, *Chem. Rev.* **112**, 75 (2012).
- ⁴⁰M. Torheyden and E. F. Valeev, *J. Chem. Phys.* **131**, 171103 (2009).
- ⁴¹H.-J. Werner and P. J. Knowles, *J. Chem. Phys.* **82**, 5053 (1985).
- ⁴²T. Yanai, Y. Kurashige, D. Ghosh, and G. K.-L. Chan, *Int. J. Quantum Chem.* **109**, 2178 (2009).
- ⁴³G. Knizia and G. K.-L. Chan, *Phys. Rev. Lett.* **109**, 186404 (2012).
- ⁴⁴G. Knizia and G. K.-L. Chan, *J. Chem. Theory Comput.* **9**, 1428 (2013).
- ⁴⁵Q. Chen, G. H. Booth, S. Sharma, G. Knizia, and G. K.-L. Chan, *Phys. Rev. B* **89**, 165134 (2014).
- ⁴⁶A. J. Coleman, *Rev. Mod. Phys.* **35**, 668 (1963).
- ⁴⁷C. Garrod and J. K. Percus, *J. Math. Phys.* **5**, 1756 (1964).
- ⁴⁸Z. Zhao, B. J. Braams, M. Fukuda, M. L. Overton, and J. K. Percus, *J. Chem. Phys.* **120**, 2095 (2004).
- ⁴⁹N. S. Blunt, T. W. Rogers, J. S. Spencer, and W. M. C. Foulkes, *Phys. Rev. B* **89**, 245124 (2014).
- ⁵⁰N. S. Blunt, A. Alavi, and G. H. Booth, (2014), arXiv:1409.2420.
- ⁵¹G. K.-L. Chan, M. Kállay, and J. Gauss, *J. Chem. Phys.* **121**, 6110 (2004).
- ⁵²H. Flyvbjerg and H. G. Petersen, *J. Chem. Phys.* **91**, 461 (1989).
- ⁵³L. Kong and E. F. Valeev, *J. Chem. Phys.* **135**, 214105 (2011).
- ⁵⁴L. Kong and E. F. Valeev, *J. Chem. Phys.* **133**, 174126 (2010).
- ⁵⁵S. Sharma, T. Yanai, G. H. Booth, C. J. Umrigar, and G. K.-L. Chan, *J. Chem. Phys.* **140**, 104112 (2014).
- ⁵⁶It should be noted that there is also potential for a small bias in the wavefunction sampling (which is also systematically improvable with increasing walker number) due to the fluctuation in the E_S parameters and consequent changes in walker population. However, this error is generally only $\mathcal{O}[10^{-5}]E_h$, and so can be considered negligible for these system⁵⁷.
- ⁵⁷W. A. Vigor, J. S. Spencer, M. J. Bearpark, and A. J. W. Thom, (2014), arXiv:1407.1753.
- ⁵⁸R. E. Thomas, C. Overy, G. H. Booth, and A. Alavi, *J. Chem. Theory Comput.* **10**, 1915 (2014).
- ⁵⁹G. H. Booth and G. K.-L. Chan, *J. Chem. Phys.* **137**, 191102 (2012).
- ⁶⁰S. Ten-no, *J. Chem. Phys.* **138**, 164126 (2013).