Exploiting locality in FCIQMC for fast excitation generation

Oskar Weser,∗,† Ali Alavi,†,‡ and Giovanni Li Manni∗,†

†Max-Planck-Institute for Solid State Research, Stuttgart, Germany
‡Yusuf Hamied Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK

E-mail: oskar.weser@gmail.com; g.limanni@fkf.mpg.de

Abstract

In this paper we propose an improved excitation generation algorithm for the full configuration interaction quantum monte carlo (FCIQMC) method, which is particularly effective in systems described by localized orbitals. The method is an extension of the precomputed heat-bath (PCHB) strategy of Holmes et al., with more effective sampling of double excitations and a novel approach for non-uniform sampling of single excitations. We demonstrate the effectiveness of the algorithm for a chain of 30 hydrogen atoms with atom-localized orbitals, a stack of benzene molecules, an Fe-porphyrin model complex, whereby we show an overall efficiency gain by a factor of two to four, as measured by variance reduction per wall-clock time.

1 Introduction

The full configuration interaction quantum monte carlo (FCIQMC) method is a sparse and highly parallelisable CI eigensolver in quantum chemistry.2,3 Relying on the sparsity of the
CI Hamiltonian and its eigensolutions, FCIQMC is in general not limited by the active space sizes,\textsuperscript{4-12} in contrast to conventional CI methods,\textsuperscript{13,14} which are limited to at most 18 electrons and 18 orbitals.\textsuperscript{15} Due to the nearly linear parallelization of the algorithm, FCIQMC takes full advantage of modern computer architectures,\textsuperscript{3} and allows a large number of electrons and orbitals to be explicitly correlated; active spaces containing up to 96 electrons in 159 orbitals have been reported to date.\textsuperscript{10} The FCIQMC dynamics also allows to stochastically sample reduced density matrices (RDMs) via the replica trick,\textsuperscript{16} enabling the calculation of properties and orbital optimization, in the form of the stochastic complete and generalized active space self consistent field (Stochastic-CASSCF and Stochastic-GASSCF, respectively) methods.\textsuperscript{4,10}

In FCIQMC, the stochastically sampled wave function is represented by walkers distributed across the electronic configurations that form the CI vector space. The CI coefficient of a given configuration is proportional to the expectation value of the walker occupation on that configuration averaged over time, $c_i \propto \langle N^w_i \rangle_\tau$. In the limit of large walker numbers, the approximated $c_i$ values become exact. Electronic configurations corresponding to vanishingly small $c_i$ coefficients are never or only rarely occupied by walkers. Since only occupied configurations have an impact on the computational costs of a FCIQMC simulation, the method relies on and takes advantage of sparsity. The initiator approximation (i-FCIQMC)\textsuperscript{17-19} and its further refinement, the adaptive shift method,\textsuperscript{20,21} increases the sparsity of the wave function and greatly improve the convergence with respect to the walker number.

The FCIQMC wave function is propagated along the imaginary time, $\tau$, in discretized steps, $\Delta \tau$. The iterative procedure of the FCIQMC dynamics consists of four main stages: excitation generation, spawn, death, and annihilation. In the excitation generation stage, new configurations $|D_j\rangle$ are suggested as targets. In the spawn step, new walkers are then spawned to $|D_j\rangle$ from $|D_i\rangle$ ($i \neq j$), with a probability $p_{\text{spawn}} \propto \Delta \tau \langle D_i | \hat{H} | D_j \rangle$. In the death step, configurations $|D_i\rangle$ are killed with a probability proportional to their diagonal matrix element $\propto \Delta \tau \langle D_i | H | D_i \rangle$. In the annihilation step, walker contributions (arising
from different spawning events) residing on the same configuration are summed up with their respective signs. If they are of identical amplitude and opposite sign, they annihilate each other and any information about that configuration is removed from the data tables.

The exact details of the four main stages of the FCIQMC algorithm greatly affect its performance. For example, a low rate of successful spawns makes the algorithm inefficient; multiple spawns from one configuration to another (a “bloom” event) cause the calculation to become unstable. In order to reduce inefficiencies, new configurations \( |D_j \rangle \) should not be suggested uniformly in the excitation generation step. Instead, they should be selected with a probability \( p_{\text{gen}} \propto \langle D_i | \hat{H} | D_j \rangle \), such that important connections are sampled more often and the spawn probability \( p_{\text{spawn}} \propto \Delta \tau \frac{\langle D_i | \hat{H} | D_j \rangle}{p_{\text{gen}}} \) becomes nearly constant. This non-uniform (often also referred to as weighted) excitation generation is at the heart of an efficient FCIQMC algorithm.

For double excitations several algorithms exist with different trade offs between efficiency and memory demand, notable ones include: (a) the Cauchy-Schwarz excitation generator by the Alavi group,\(^3\) (b) the heat-bath Power-Pitzer method\(^{22}\) by Neufeld and Thom, and (c) the precomputed heat-bath (PCHB) method\(^1\) by Holmes et al.. PCHB is the foundation of the work presented in this document.

One important property that these fast double excitation generators rely on, is that the magnitude of the matrix element between two determinants, connected by a double excitation, does not depend on the involved determinants (details in section 2). This property allows the determinant-independent precalculation of probability distributions and their efficient sampling.\(^1,3,22,23\)

In PCHB the sampling of double excitations \((AB \leftrightarrow IJ)\) occurs in multiple steps. In the particle selection, a first particle \( I \) is chosen according to probability \( p(I) \), then a second particle \( J \) is chosen under the conditional probability \( p(J|I) \). Next, in the hole selection, holes \( A \) and \( B \) are chosen with probabilities \( p(A|IJ) \) and \( p(B|IJA) \), respectively. When suitably localized MOs are used, after particle \( I \) has been selected, the second particle has
a high probability, \( p(J|I) \), to be selected from the vicinity of \( I \), rather than from spatially distant orbitals. Hence the PCHB algorithm can take advantage of locality effects.

Furthermore, particle indices, \( I \) and \( J \), should preferably be guaranteed to be occupied in the starting determinant, \( |D_i \rangle \), while the hole indices \( A \) and \( B \) should be guaranteed to be empty, which requires additional work. This implies that there is a trade off between (a) faster sampling, which cannot guarantee occupied \( I, J \) and empty \( A, B \), versus (b) slower but high quality sampling, that can guarantee this. If the underlying probabilities are near-uniform, then the advantage of weighted sampling vanishes and it might be beneficial to return back to uniform sampling, which is faster and cheaply guarantees occupied \( I, J \) and empty \( A, B \).

The Hamiltonian matrix elements for single excitations, on the other hand, depend on the occupied orbitals hence on the determinant from which the excitation takes place (details in section 2). This determinant-dependency makes the precalculation of single excitation probabilities harder, and single excitations have usually been sampled uniformly.\(^{1,3,10,22}\) In order to improve upon the uniform single excitation generation, Neufeld and Thom suggested to build precalculated probabilities for single excitations relying on the reference determinant, with some modifications.\(^{22}\) This approach works best for systems featuring single-reference wave functions. In the context of variational Monte Carlo, Sabzevari and Sharma suggested to contract over all orbitals in the two-electron contribution to the single excitation Hamiltonian matrix element, and to truncate all single excitations that are below a user-defined threshold.\(^{24}\) This contraction over all orbitals is independent from the determinants and can be precomputed, but the truncation introduces an artificial bias in the excitation generation; moreover, it is still uniform for all those excitations that are above the threshold.

For delocalized molecular orbital bases, double excitations are usually more important and numerous than single excitations. Single excitations often feature vanishingly small matrix elements, for they (partially) fulfill the Brillouin theorem in its single-reference or generalized form.\(^{25}\) Therefore, most effort in the community has gone into improving the dou-
ble excitation generation. However, in localized MO bases single excitations become as (or more) important than double excitations, capturing mean field effects via the CI-expansion. The corresponding single excitation Hamiltonian matrix elements are then generally large and pivotal in the wave function optimization. Thus, the development of efficient sampling schemes for single excitations becomes particularly relevant for localized orbital bases. These orbital bases play a crucial role in our recently discovered compression of spin-adapted wave functions of ground and excited states of exchange coupled polynuclear transition metal clusters. 6–8,12,26–28

Inspired and motivated by our research carried on localized orbital bases, we have developed a novel PCHB algorithm that, relying on locality of the correlation effects, has an increased excitation generation efficiency for both single and double excitations. This algorithm is the focus of this work.

For our new precomputed and weighted single excitation generator, we contract over all orbitals to obtain an approximated estimate for the non-uniform sampling probabilities of single excitations. The contraction is similar to the work of Sabzevari and Sharma but we do not introduce a bias from truncation, sample the single excitations non-uniformly and show that the non-uniform probabilities are well approximated in localized molecular orbital bases.

For the double excitations we systematically investigated the tradeoff between speed and guaranteeing occupied or empty orbitals in PCHB and determined which combination of uniform or (constrained) weighted PCHB sampling is the fastest, to date not present in the literature. In this work we show that the choice matters, considerably improves the excitation generation of doubles, and depends on the system at hand. We also improve upon the current literature by examining how to efficiently draw from constrained distributions, for example $p(I)_{I \in |D\rangle}$ from the precomputed $p(I)$.

The new PCHB algorithm for singles and the improved excitation generation for doubles allows FCIQMC dynamics that are 2 to 4 times more efficient than the implementations
developed by Holmes et al.\textsuperscript{1} and by us (within the \textsc{nci}\textsuperscript{3} project).

The novel PCHB algorithm is easily extended to the recently developed Stochastic-GAS wave function optimizations\textsuperscript{10} and to the spin-purification strategy in Slater-determinant based FCIQMC.\textsuperscript{11}

The remainder of the article is organized as follows: In section 2, we give the necessary background for our novel PCHB excitation generator in FCIQMC. This includes, the equations for the evaluation of Hamiltonian matrix elements, their decay with increasing spatial distance, and the already existing PCHB algorithm for double excitations; in section 3, we introduce our improved sampling of double excitations; in section 4, we introduce our new weighted, precomputed single excitation generation; in section 5, we demonstrate the application of the novel PCHB on several test case molecular systems and quantify the efficiency improvements: A chain of 30 hydrogen atoms with atom-localized orbitals, whose Hamiltonian is dominated by single excitation couplings, demonstrates the capabilities of our new algorithm for sampling single excitations; stacks of benzene molecules, with fragment-localized MOs serves as example for a system characterized by strong correlation effects (double excitations) within each fragment, and generally small correlation effects across the fragments; the efficiency of the new PCHB is numerically demonstrated also for highly delocalized MOs, using an Fe-porphyrin model system from previous works,\textsuperscript{4,5,9,10} as well as the \textit{N}\textsubscript{2} dimer at equilibrium geometry with a large full-CI space. In sections 6 and 7, we discuss how both generalized active spaces and spin purification in a Slater determinant basis benefit from the improved PCHB algorithm. Section 8 collects our final remarks and conclusions of the work.

2 Theoretical background

This section describes the key elements that represent the foundation of the novel PCHB algorithm suggested in the present work.
2.1 Matrix elements

Using the conventions of the “purple book”\textsuperscript{25} and assuming orthonormal spin orbitals we write the second-quantized, non-relativistic molecular Hamiltonian operator in the Born-Oppenheimer approximation with atomic units as

\begin{equation}
\hat{H} = \sum_{QP} h_{PQ} a_{Q}^{\dagger} a_{P} + \frac{1}{2} \sum_{RQSP} g_{RPSQ} a_{R}^{\dagger} a_{S}^{\dagger} a_{Q} a_{P} + E_{\text{nuc}} \tag{1}
\end{equation}

where \( P, Q, R, S \) are generic spin orbital indices. The \( a^{\dagger} \) and \( a \) are the usual unitary, second-quantized creation and annihilation operators. The one-electron integral \( h_{AI} \) is given as

\begin{equation}
h_{AI} = \int \phi_{A}^{*}(x) \left( -\frac{1}{2} \nabla^{2} - \sum_{X \in \text{nuclei}} \frac{Z_{X}}{|R_{X} - r|} \right) \phi_{I}(x) \, dx \tag{2}
\end{equation}

where \( \phi_{I} \) denotes the \( I \)-th spin orbital, \( r \) and \( x \) refer to the spatial and the combined spatial and spin electronic coordinate, respectively, and \( Z_{X} \) and \( R_{X} \) are the charge and spatial coordinate of the \( X \)-th nucleus. The two-electron integral \( g_{ABIJ} \) is given in chemist’s notation as

\begin{equation}
g_{ABIJ} = \int \int \phi_{A}^{*}(x_{1}) \phi_{B}^{*}(x_{2}) \frac{1}{r_{12}} \phi_{I}(x_{1}) \phi_{J}(x_{2}) \, dx_{1} \, dx_{2} \tag{3}
\end{equation}

where \( r_{12} = |r_{1} - r_{2}| \). The nuclear repulsion term, \( E_{\text{nuc}} \), is the classical repulsion of point charges given by

\begin{equation}
E_{\text{nuc}} = \sum_{X < Y; X, Y \in \text{nuclei}} \frac{Z_{X} Z_{Y}}{|R_{X} - R_{Y}|} \tag{4}
\end{equation}

The Hamiltonian operator contains up to two-particle excitation terms and in general the matrix element \( \langle D_{i} | \hat{H} | D_{j} \rangle \) between two determinants is zero if they differ by occupation in more than two spin-orbitals (in virtue of the Slater-Condon rules\textsuperscript{29}). The nonzero terms
are given as:

\[ \hat{H}\ket{D_i}^{\text{(diag)}} := \bra{D_i} \hat{H} \ket{D_i} = \sum_{I \in |D_i\rangle} h_{II} + \frac{1}{2} \sum_{I,J \in |D_i\rangle} (g_{IIJJ} - g_{IJJI}) \]  

(5)

\[ \hat{H}\ket{D_i}^{A \leftrightarrow I} := \bra{D_i} \hat{H} \ket{a^\dagger_A a^I D_i} = \Gamma_{AI} D_i h_{AI} + \frac{1}{2} \sum_{R \in |D_i\rangle} (g_{AIRR} - g_{ARRI}) \]  

(6)

\[ \hat{H}\ket{D_i}^{AB \leftrightarrow IJ} := \bra{D_i} \hat{H} \ket{a^\dagger_A a^\dagger_B a^I a^J D_i} = \Gamma_{AI} \Gamma_{BJ} D_i (g_{AIBJ} - g_{AJBI}) \]  

(7)

where indices \( I < J \) and \( A < B \) refer to occupied and unoccupied spin-orbitals in \(|D_i\rangle\), respectively, connecting to \(|D_j\rangle\). We write \( I \in |D_i\rangle \) if \( I \) is occupied in determinant \(|D_i\rangle\), hence \( R \) in eq. 6 represents an index running over all the occupied spin-orbitals. The fermionic prefactor is \( \Gamma_{AI} D_i = (-1)^n \), where \( n \) is the number of occupied spin-orbitals between \( A \) and \( I \) in determinant \(|D_i\rangle\). Note also, that the indices for the single and double excitations have to be pairwise different (\( I \neq J, A \neq B, \{I,J\} \cap \{A,B\} = \emptyset \)). We will write uppercase \( I \) for spin-orbitals, lowercase \( i \) for spatial orbitals, and \( i_\sigma \) for the \( i \)-th spatial orbital with spin \( \sigma \).

Our new PCHB excitation generator relies on the decay of Hamiltonian matrix elements from vanishing integrals. We assume the MOs to be suitably localized and separated enough from each other so that a distance \( R \) between the MOs can be defined. We have to emphasize that the following discussion does not apply to delocalized orbitals. For the one-electron integral there exists a constant \( k \), such that

\[ |h_{IA}| \leq k \cdot S_{IA} \]  

(8)

where \( S_{IA} = \bra{\phi_I} \phi_A \rangle \) is the overlap between the two spin orbitals which decays as \( O(\exp(-R^2)) \) with the distance \( R \) between the MOs.\(^{30}\) An intuitive understanding comes from the observation that with less and less overlap the two functions \( \phi_I \) and \( \phi_A \) have more and more disjoint support with respect to the integration variable (see equation 2).
The two-electron integrals decay\textsuperscript{31} for large distances with a constant $k$ as

$$|g_{AIBJ}| \leq k \cdot \frac{S_{AI}S_{BJ}}{R}.$$  \hspace{1cm} \text{(9)}

This means we can qualitatively distinguish two cases: the classical Coulomb term of the type $g_{IIJJ}$ and all other non-classical integrals. In the case of the Coulomb term we have $S_{II} = S_{JJ} = 1$ which implies

$$|g_{IIJJ}| \leq k \cdot \frac{1}{R}.$$  \hspace{1cm} \text{(10)}

This decay can be related to classical electrostatics, since the orbitals \textit{depending on the same electronic coordinate} can be combined into the classical electron density $\rho_I = \phi_I^* \phi_I$ which gives

$$g_{IIJJ} = \int \int \rho_I(x_1) \frac{1}{r_{12}} \rho_J(x_2) \, dx_1 \, dx_2 = \int \int \rho_I(r_1) \frac{1}{r_{12}} \rho_J(r_2) \, dr_1 \, dr_2.$$  \hspace{1cm} \text{(11)}

and decays, like classical point charges, with $O\left(\frac{1}{R}\right)$ for a large enough distance $R$ between the MOs $\phi_I$ and $\phi_J$. The non-classical two-electron integrals are instead dominated by the exponential decay of the overlap terms, $S_{AI}$ and $S_{BJ}$, and decay much faster than $1/R$.

As shown in equation 5, Coulomb interactions appear only in the diagonal of the Hamiltonian; thus, we expect an exponential decay with $O(\exp(-R^2))$ for the off-diagonal Hamiltonian matrix elements.

\section{2.2 Excitation Probabilities}

The aim of excitation generators in FCIQMC is to suggest a new determinant $|D_j\rangle$, from $|D_i\rangle$, with a probability proportional to their matrix element. So the optimal excitation generator would sample with the following probabilities:

$$p_{\text{gen}}(\langle D_j \rangle \leftarrow \langle D_i \rangle) = \frac{|\langle D_i | \hat{H} | D_j \rangle|}{\sum_{|D_x\rangle \sim |D_i\rangle} |\langle D_i | \hat{H} | D_x \rangle|}.$$  \hspace{1cm} \text{(12)}

9
Here and in the following we write $|D_i⟩ \sim |D_j⟩$ if two determinants are connected, and $|D_j⟩ \leftarrow |D_i⟩$ for a specific excitation from $|D_i⟩$ to $|D_j⟩$. Moreover, the labels $|D_i⟩ \overset{1}{\sim} |D_j⟩$ or $|D_i⟩ \overset{2}{\sim} |D_j⟩$ are used to indicate whether the two determinants are connected via single or double excitations, respectively, and $|D_j⟩ \overset{1}{\leftarrow} |D_i⟩$ or $|D_j⟩ \overset{2}{\leftarrow} |D_i⟩$ to indicate whether $|D_j⟩$ is obtained via single or double excitations from $|D_i⟩$ to $|D_j⟩$, respectively.

We first decide with probabilities $p_1$ and $p_2$ if we perform a single or double excitation respectively and then sample from the conditional probabilities $p_{\text{gen}}(|D_j⟩ \leftarrow |D_i⟩|n), n \in \{1, 2\}$. In order to maximize the possible time step $\Delta \tau$ while keeping a stable calculation, the probabilities $p_1, p_2$ are adjusted during the calculation to preserve

$$p_1 \cdot \min_{\overset{\sim}{|D_i⟩}} \left( \frac{p_{\text{gen}}(|D_j⟩ \leftarrow |D_i⟩|1)}{\langle D_i | \hat{H} | D_j \rangle} \right) = p_2 \cdot \min_{\overset{\leftarrow}{|D_i⟩}} \left( \frac{p_{\text{gen}}(|D_j⟩ \leftarrow |D_i⟩|2)}{\langle D_i | \hat{H} | D_j \rangle} \right). \quad (13)$$

From equation 13 it becomes apparent that if an excitation generator is of low quality, i.e. there is a pair $|D_i⟩ \overset{\sim}{\sim} |D_j⟩$ with a low ratio $\frac{p_{\text{gen}}(|D_j⟩ \leftarrow |D_i⟩|n)}{\langle D_i | \hat{H} | D_j \rangle}$, then the corresponding $p_n$ become large. For example, if single excitations are sub-optimal they will be sampled more often based on eq. 13.

Localized MOs tend to have considerably large matrix elements for single excitations and these matrix elements are non-uniform, because excitations to distant MOs are negligible (vide infra). The generally utilized uniform excitation generation for single excitations, in the case of localized orbital bases, would lead to too small $p_{\text{gen}}$ for single excitations, due to the numerous vanishingly small single connections among distant MOs, leading to exaggeratedly large $p_1$ values, and overshadow any improvement of PCHB for double excitations. In these cases, the uniform excitation generator for single excitations represents a particular poor choice, and would negatively impact the entire dynamics. The present work overcome such limitation, and provides more balanced excitation generations.
2.3 Original PCHB strategy

In this section we summarize the key elements of the original work on PCHB by Holmes and co-workers. The exact weighted sampling of orbitals for double excitations from \( |D_i\rangle \) is given by

\[
p_{|D_i\rangle}(AB \leftarrow IJ) = \frac{|\hat{H}_{|D_i\rangle}(AB \leftarrow IJ)|}{\sum_{KL \in |D_i\rangle, CD \notin |D_i\rangle} |\hat{H}_{|D_i\rangle}(CD \leftarrow KL)|}
\]  

(14)

Clearly, whether the element is zero or non-zero depends on occupation (\( I, J \in |D_i\rangle \) and \( A, B \notin |D_i\rangle \)), hence on the determinant. When non-zero, double excitations have the important property that the magnitude is independent from the starting determinant \( |D_i\rangle \) (compare equation 7). If we write for the weight

\[
W_{IJ}^{AB} = \begin{cases} 
|g_{AIBJ} - g_{AJBI}| & (I \neq J) \land (A \neq B) \land \{I, J\} \cap \{A, B\} = \emptyset \\
0 & \text{else}
\end{cases}
\]

(15)

then equation 14 can be factorized without approximations, as

\[
p_{|D_i\rangle}(AB \leftarrow IJ) = \frac{\sum_{L \in |D_i\rangle, CD \notin |D_i\rangle} W_{IL}^{CD} \cdot \sum_{CD \notin |D_i\rangle} W_{IJ}^{CD}}{p_2^{\text{exact}}(I)} \cdot \frac{\sum_{CD \notin |D_i\rangle} W_{IJ}^{AD}}{p_2^{\text{exact}}(A|IJ)} \\
\cdot \frac{\sum_{D \notin |D_i\rangle} W_{IJ}^{AD}}{p_2^{\text{exact}}(B|IJ,A)} \cdot \frac{\sum_{LC \notin |D_i\rangle} W_{IL}^{CD}}{p_2^{\text{exact}}(J|I)} \cdot \frac{\sum_{CD \notin |D_i\rangle} W_{IJ}^{CD}}{p_2^{\text{exact}}(B|IJA)}
\]

(16)

The crucial approximation of the original PCHB is the contraction over all indices and hence the determinant-independent precalculation of probabilities which are given by:

\[
p_2^{\text{PCHB}}(I) = \frac{\sum_{LCD} W_{IL}^{CD}}{\sum_{KLCD} W_{KL}^{CD}} \quad p_2^{\text{PCHB}}(J|I) = \frac{\sum_{CD} W_{IJ}^{CD}}{\sum_{LCD} W_{IL}^{CD}} \\
p_2^{\text{PCHB}}(A|IJ) = \frac{\sum_{AD} W_{IJ}^{AD}}{\sum_{CD} W_{IJ}^{CD}} \quad p_2^{\text{PCHB}}(B|IJA) = \frac{W_{IJ}^{AB}}{\sum_{D} W_{IJ}^{AD}}
\]

(17)
Considering that the order in which we pick the particles and holes does not matter for the final determinant, and that in general one cannot assume symmetry, e.g. \( p(I|J) \neq p(J|I) \), the full probability of selecting \(|D_j⟩ = a_1^† a_2^† a_J a_I |D_i⟩\) can then be calculated via:

\[
p_{|D_i⟩}(\{AB\} \leftarrow \{IJ\}) = p_2 \cdot (p_2^{\text{PCHB}}(I) \cdot p_2^{\text{PCHB}}(J|I) + p_2^{\text{PCHB}}(J) \cdot p_2^{\text{PCHB}}(I|J)) \\
\cdot (p_2^{\text{PCHB}}(A|IJ) \cdot p_2^{\text{PCHB}}(B|IJA) + p_2^{\text{PCHB}}(B|IJ) \cdot p_2^{\text{PCHB}}(A|IJB))
\]

(18)

The precomputed probabilities from equation 17 have the huge advantage that they can be calculated once at the beginning of the calculation. In combination with the alias sampling algorithm\(^{23,32}\) the precalculated probabilities allow sampling with \( O(1) \) run time cost. This is a typical trade off between memory and time. The memory requirements can in part be reduced by using read-only, node-shared memory for storing the alias tables.\(^3,10\)

It would be desirable for an excitation generator that particle indices, \( I, J \), and hole indices, \( A, B \), are guaranteed to be occupied and unoccupied in \(|D_i⟩\), respectively. Unfortunately, sampling according to equation 17, with no reference to the current determinant, \(|D_i⟩\), has the important disadvantage that the particle/hole indices suggested might not be occupied/empty in \(|D_i⟩\); this excitation can only be discarded.

### 3 Improved PCHB for double excitations

Our improvements to the sampling of doubles in PCHB are about when and how to guarantee if orbitals are occupied or empty. The necessary renormalizations for constrained sampling (a sampling where the notion of occupied/empty orbital is available) that guarantees \( I, J \) and \( A, B \) to be occupied and empty, respectively, in \(|D_i⟩\), read as:

\[
P(I)_{|I\in|D_i⟩} = \frac{p(I)}{\sum_{X\in|D_i⟩} p(X)} \quad p(J)_{|J\in|D_i⟩} = \frac{p(J)}{\sum_{X\in|D_i⟩} p(X)} \\
p(A|IJ)_{|A\notin|D_i⟩} = \frac{p(A|IJ)}{\sum_{X\notin|D_i⟩} p(X|IJ)} \quad p(B|IJA)_{|B\notin|D_i⟩} = \frac{p(B|IJA)}{\sum_{X\notin|D_i⟩} p(X|IJA)}
\]

(19)
These probabilities scale with $O(N_e)$ and $O(n_{orb} - N_e)$ for $I, J$ and $A, B$, respectively.

A clear trade off emerges from the comparison of eq. 17 and eq. 19, between (a) drawing from precomputed distributions with constant scaling and subsequently discarded excitations, and (b) from ad hoc constrained probability distributions with linear scaling and no discarded excitations. We will refer to unconstrained sampling $(p(I), \ldots)$ as fast-weighted (eq. 17), and to constrained sampling $(p(I)|_{I \in |D_i\rangle}, \ldots)$ as fully-weighted sampling (eq. 19). The uniform sampling, also used for comparison in this work, allows to draw from constrained subsets with $O(1)$ time. This implies that for near-uniform probability distributions the uniform sampling might outperform weighted sampling, after taking quality of excitation and iteration time into account. The possible sampling schemes are summarized in Table 1 and, in principle, can be freely combined among each other for the selection of first/second particle or hole. Note that fast-weighted and fully-weighted methods have additional, non-negligible memory and one-time initialization cost, that scale with $O(n_{orb}^1)$, $O(n_{orb}^2)$, $O(n_{orb}^3)$, and $O(n_{orb}^4)$ for the first, second particle and first, second hole, respectively. We will see later

![Table 1: Different Sampling Methods and their run time scaling.](image)

that the specific choice of combinations has large implications for the performance of the overall FCIQMC dynamics. In particular fast-weighted sampling works best for the hole-selection of active spaces with an excess of unoccupied orbitals, i.e. if there are many empty orbitals then $p_2^{PCHB}(A|IJ) \approx p_2^{PCHB}(A|IJ)|_{A \notin |D_i\rangle}$, and performing the full weighting would only result in additional, unnecessary computational effort. Similar conclusions would be reached for active space with an excess of occupied orbitals, but those are more rare in prac-
tical applications and limited by the Pauli principle and will not be discussed any further. Uniform sampling works best for active spaces with near-uniform excitation probabilities and has no additional costs to guarantee orbitals to be occupied or unoccupied. Fully-weighted sampling is the slowest in wall-clock time, but samples with the highest quality.

To date no systematic investigation is available, on which combination of uniform or (constrained) weighted PCHB sampling is the fastest. The original PCHB paper\(^1\) and the work by Neufeld and Thom\(^2\) assumed fully-weighted particle selection \((p^{\text{PCHB}}_2(I)|I\in|D_i\rangle\) and \(p^{\text{PCHB}}_2(J|I)|J\in|D_i\rangle\) ) and fast-weighted hole selection \((p^{\text{PCHB}}_2(A|IJ)\) and \(p^{\text{PCHB}}_2(B|IJ\!A))\) to be the fastest combination. This choice was likely motivated by the systems investigated in their work, where FCI calculations of small molecular systems were carried to the basis set limit. This implies a fixed number of particles \(N_e\) and an ever growing number of empty spin orbitals, \(n_{\text{orb}}\). In our earlier work using PCHB\(^3,10\) we have assumed uniform particle selection \((p_{\text{uni}}^2(IJ)|I,J\in|D_i\rangle\) ) and fast-weighted hole selection \((p^{\text{PCHB}}_2(AB|IJ))\) to be the fastest combination. Note that directly sampling pairs \(IJ\) and \(AB\) can reduce the memory cost as described in reference 3, but is tied to a fast-weighted sampling scheme.

Based on the previous discussion, our improvements to the sampling of double excitations in PCHB are twofold: (a) we suggest optimal combinations of fast-, fully-weighted, and uniform sampling depending on general features of the systems investigated, and (b) we improve the sampling speed from constrained distributions. We will present these improvements separately in the next two subsections.

### 3.1 Improved sampling combinations

The choice of better combinations of sampling schemes greatly depends on the probability distributions from equation 17 which are highly system dependent. Different systems may prefer different excitation generation schemes, based on the form of electronic interactions, the orbital basis of choice (delocalized or localized), and whether there is an excess of empty orbitals. We will analyze the shape of probability distributions for some selected systems,
and numerically prove our findings in the application section 5.

The first example we consider is a stack of ten benzene molecules at an inter-molecular distance of 3.0 Å, and orbitals that are fragment-localized, i.e. they are natural, Hückel-like orbitals on each benzene. The CAS(60,60) was ten times the minimal active space of six electrons in six \( \pi \)-orbitals for one benzene unit. As we showed in a previous publication,\(^\text{10}\) already single excitations between the fragments are enough to recover Full CI energy, while the system as a whole is dominated by double excitations inside each fragment.

The probability \( p_{\text{PCHB}}^2(I) \) is displayed in Figure 1. It shows that the fragments at the border have slightly different behavior than the fragments in “bulk” and there is a slight variation in the probability. Also the symmetry of repeating fragments is visible in the probabilities. Overall it is a near-uniform distribution with a relative standard deviation of 13\% from the uniform distribution. For this reason, \( p_{\text{uni}}^2(I)|_{I \in |D_i\rangle} \) is likely the best choice for the first particle, cheaply guaranteeing \( I \) to be occupied.

![Figure 1: The PCHB probability of the first particle \( p_{\text{PCHB}}^2(I) \) for a double excitation in a stack of ten benzene molecules. The orbitals are ordered by fragment and the vertical dotted lines separate MOs of different benzene molecules. Since \( \alpha \) and \( \beta \) electrons have the same probability, assuming RHF-type orbitals, we could use the spatial orbital index. The horizontal dashed line gives the probability of the corresponding unconstrained uniform distribution \( p_{\text{uni}}^2(I) \) as reference. The value is obtained as \( 1/n_{\text{orb}} = 1/120 = 8.3 \cdot 10^{-3} \).](image)
The conditional probability \( p_{2}^{\text{PCHB}}(J|I) \) to sample the second particle is displayed in Figure 2. The left, non-logarithmic plot (Figure 2a) shows that the correlation is highly local. The second particle has the highest probability to be selected on the same fragment, while quickly decaying for neighboring fragments. Any choice for the second particle further away than 6 Å (second nearest fragment) is nearly negligible. The logarithmic plot (Figure 2b) reveals the exponential decay with increasing distance. This observation shows that the distribution is highly non-uniform and weighted sampling is preferred. Moreover, as the orbital space considered, CAS(60,60) is far from a case with an excess of occupied orbitals, it is advantageous to guarantee that the second picked orbital index is indeed occupied. By this argument fully-weighted sampling with \( p_{2}^{\text{PCHB}}(J|I) |_{J \in |D_{i}\rangle} \) is recommended. In addition the weighted sampling automatically prefers geminal, i.e. opposite spin pair, excitations as can be seen from comparing the blue and orange line in Figure 2.

Figure 2: The PCHB probability of the second particle after having picked an \( \alpha \) electron in the first spatial orbital \( p_{2}^{\text{PCHB}}(J|1_{\uparrow}) \) for a double excitation in a stack of ten benzene molecules. The orbitals are sorted by fragment and the vertical dotted lines separate MOs of different benzene molecules. We use again spatial orbital indices. The blue line represents an opposite spin \( p_{2}^{\text{PCHB}}(j_{\downarrow}|1_{\uparrow}) \), while the orange line represents a parallel spin \( p_{2}^{\text{PCHB}}(j_{\uparrow}|1_{\uparrow}) \) for the second electron. The right figure is a logarithmic version of the left one. The horizontal dashed line gives the probability of the corresponding unconstrained uniform distribution \( p_{2}^{\text{uni}}(J|1_{\uparrow}) \) as reference. The value is obtained as \( 1/n_{\text{orb}} = 1/120 = 8.3 \cdot 10^{-3} \).

The subsequent hole-selections are also highly non-uniform with no excess of empty
orbits which suggests

$$p_{D_i}(AB \leftarrow IJ) = p_{D_i}^{\text{uni}}(I) |_{I \in |D_i\rangle} \cdot p_{D_i}^{\text{PCHB}}(J|I) |_{J \in |D_i\rangle} \cdot p_{D_i}^{\text{PCHB}}(A|I) |_{A \in |D_i\rangle} \cdot p_{D_i}^{\text{PCHB}}(B|JA) |_{B \in |D_i\rangle}$$

(20)

to be the best choice for such localized systems, namely uniform selection for the first particle
and fully-weighted selection for the second particle and the two holes. Notably this sampling
scheme is neither the one suggested by Holmes et al.\textsuperscript{1} nor the one we have been using in
NECI.\textsuperscript{3}

The second system is Full CI on the $N_2$ dimer at equilibrium distance with a Dunning’s
cc-pVQZ basis\textsuperscript{33} using Hartree-Fock orbitals, resulting in a (14,110) active space. This is
a single-reference system with highly delocalized orbitals that fulfill the Brioullin theorem,
which makes the double excitations particularly important.

The probability for the first particle $p_2^{\text{PCHB}}(I)$ is displayed in Figure 3. The PCHB distri-
bution exhibits a relatively small standard deviation of 24\% from the uniform distribution.
For this reason, analogously to the benzene stack, $p_2^{\text{uni}}(I) |_{I \in |D_i\rangle}$ is to be considered the best
choice for the first particle.

The conditional probability $p_2^{\text{PCHB}}(J|I)$ to sample the second particle is displayed in
Figure 4. The left, non-logarithmic plot (Figure 4a) is non-uniform; however, the range of
PCHB values is much smaller than for the stack of benzene (Figure 2a), indicating a more
homogeneous distribution. The logarithmic plot (Figure 4b) reveals that the probabilities
range mostly over “just one” order of magnitude, unlike the four orders of magnitude in
Figure 2b. There is no obvious exponential decay for the probability of the second particle,
due to the compact nature of the system. Overall, the $N_2$ dimer and similar systems might
still benefit from uniform sampling of the second particle $p_2^{\text{uni}}(J|I) |_{J \in |D_i\rangle}$. This suggestion is
investigated in greater detail and confirmed in section 5.

The subsequent hole-selections are again highly non-uniform, which suggests weighted
sampling. Due to the excess of empty orbitals the fast-weighted sampling of hole indices, is
Figure 3: The PCHB probability of the first particle $p_2^{\text{PCHB}}(I)$ for a double excitation in the N$_2$ dimer. The orbitals are sorted by probability. Since $\alpha$ and $\beta$ electrons have the same probability, assuming RHF-type orbitals, we could use the spatial orbital index. The horizontal dashed line gives the probability of the corresponding unconstrained uniform distribution $p_2^{\text{uni}}(I)$ as reference. The value is obtained as $1/n_{\text{orb}} = 1/220 = 4.5 \cdot 10^{-3}$.

Figure 4: The PCHB probability of the second particle after having picked an $\alpha$ electron in the first spatial orbital $p_2^{\text{PCHB}}(J|1\uparrow)$ for a double excitation in the N$_2$ dimer. We use again spatial orbital indices which are sorted by joint spatial probability, i.e. $p_2^{\text{PCHB}}(j_\alpha|1\uparrow) + p_2^{\text{PCHB}}(j_\beta|1\uparrow)$. The blue line represents an opposite spin $p_2^{\text{PCHB}}(j_\downarrow|1\uparrow)$, while the orange line represents a parallel spin $p_2^{\text{PCHB}}(j_\uparrow|1\uparrow)$ for the second electron. The right figure is a logarithmic version of the left one. The horizontal dashed line gives the probability of the corresponding unconstrained uniform distribution $p_2^{\text{uni}}(J|1\uparrow)$ as reference. The value is obtained as $1/n_{\text{orb}} = 1/220 = 4.5 \cdot 10^{-3}$.
expected to outperform the fully weighted strategy, since a randomly picked index is more likely to be unoccupied. The above considerations suggest the following sampling scheme

\[ p_{|D_i\rangle}(AB \leftarrow IJ) = p_2^{\text{uni}}(I)|I \in |D_i\rangle \cdot p_2^{\text{uni}}(J|I)|J \in |D_i\rangle \cdot PCHB(A|IJ) \cdot PCHB(B|IJA) \]  

(21)

for \( N_2 \) and similar compact systems featuring delocalized orbital bases and active spaces with an excess of empty orbitals. Notably this sampling scheme is the choice we have been using in NECI\(^3\) and differs from the one suggested by Holmes et al..\(^1\)

### 3.2 Faster constrained sampling

This subsection describes how to efficiently draw from constrained distributions, for example \( p(I)|I \in |D_i\rangle \) from the precomputed \( p(I) \). We assume that both the unconstrained probabilities \( p(I) \) and the corresponding alias tables for alias sampling are preconstructed, i.e. we can draw in \( \mathcal{O}(1) \) from \( p(I) \).

In the literature it has so far been assumed, that constrained distributions are best sampled by rebuilding alias tables for the constrained probabilities \( p(I)|I \in |D_i\rangle \).\(^1,2,2\) This approach still uses the precomputed probabilities \( p(I) \) to obtain \( p(I)|I \in |D_i\rangle \), but ignores the already available alias tables for \( p(I) \). This is an unnecessary slowdown because the computational effort for reconstructing constrained alias tables scales linearly with the number of particles \( \mathcal{O}(N_e) \) or the number of empty orbitals \( \mathcal{O}(n_{\text{orb}} - N_e) \) for constraints of the type \( I \in |D_i\rangle \) or \( A \notin |D_i\rangle \), respectively.

We suggest to redraw from \( p(I) \), using the already available alias tables, until the required (un-)occupied index is sampled (see Alg. 3). If \( |D_i\rangle \) is represented as a bit mask, then the test for membership \( I \in |D_i\rangle \) has \( \mathcal{O}(1) \) cost and since drawing from \( p(I) \) also has \( \mathcal{O}(1) \) cost, the whole operation of redrawing has \( \mathcal{O}(1) \) cost.

In order to obtain the renormalized probabilities (equation 19) we need equations of the type \( p(I)|I \in |D_i\rangle = p(I) / (\sum_{X \in |D_i\rangle} p(X)) \), which would scale with \( \mathcal{O}(N_e) \) because of the sum at
the denominator; for hole constraints $p(A|I)|_{A \notin D_i} = p(A|I)/(\sum X \notin D_i p(X|I))$ the scaling would be $O(n_{\text{orb}} - N_e)$. Since there is the following relationship for the complement

$$\sum_{I \in D_i} p(I) = 1 - \sum_{A \notin D_i} p(A)$$

(22)

we can choose to scale with the number of particles or empty orbitals, whichever is cheaper. This is particularly advantageous for active spaces with an excess of empty orbitals.

The redrawing method and the method of reconstructing the alias tables have both linear scaling and they require the same renormalizations; however, reconstructing alias tables implies additional work (see Alg. 2), hence, the scaling prefactor is lower for the redrawing method. The higher the weight of the constrained subset, the fewer misses there are in the redrawing method. Hence the redrawing method becomes better in comparison, if the weight of the constrained subset is higher.

Figure 5 depicts the typical situation of an active space with a similar number of particles and spatial orbitals, hence the constrained subset, for both particle- and hole-selection, contains half the elements of the total system size. In the same Figure we can see the linear scaling for both methods and that the redrawing method is roughly five times faster than rebuilding the alias sampling tables. For higher weights of the constrained subset, the redrawing method can get up to ten times faster. Further details of the timings are discussed in the Appendix A.1.

In the context of FCIQMC the faster sampling from constrained distributions is relevant for both single and double excitations; the singles are described in the next section. However it can have broader applicability to general non-uniform, discrete probability distributions even in fields outside quantum chemistry.
Figure 5: Sampling times for drawing from a constrained subsets, that contains half the elements of the total system size. The colors denote the algorithm: blue uses the redrawing method (algorithm 3), orange rebuilds an alias table for the constrained subset.

4 Novel scheme for precomputed single excitations

As discussed earlier, an accurate and balanced generation of single and double excitations is of paramount importance for efficient FCIQMC dynamics. According to eq. 13 a sub optimal single excitation generator leads to an exceedingly frequent sampling of single excitations, which overshadows any improvement for double excitation generators.

In this section we introduce an efficient way to precompute the probabilities for single excitations. As for the precomputed double excitations we start with the exact, determinant-dependent, factorized sampling probability

$$p_{|D_i⟩}(A ← I) = \frac{\sum_{C \notin |D_i⟩} |\tilde{H}_{|D_i⟩}(C ← I)|}{\sum_{K \notin |D_i⟩, C \notin |D_i⟩} |\tilde{H}_{|D_i⟩}(C ← K)|} \cdot \frac{|\tilde{H}_{|D_i⟩}(A ← I)|}{\sum_{C \notin |D_i⟩} |\tilde{H}_{|D_i⟩}(C ← I)|}$$

(23)

$$= p_1^{\text{exact}}(I) \cdot p_1^{\text{exact}}(A|I) .$$

Equation 23 depends in two ways on the starting determinant $|D_i⟩$. The first, obvious
dependence, is the restricted summation over (un-)occupied orbitals (the $K$ and $C$ indices). This dependence is also present for double excitations and is approximated in PCHB by contracting over all orbitals (see eq. 16). The additional dependence for single excitations is that the magnitude of the matrix element highly depends on the starting determinant, see eq. 6. For example, according to the Brioullin-Theorem, single excitations from the Hartree-Fock reference determinant ($D_{HF}$) should have a vanishing matrix element

$$\forall I \in D_{HF}, A \notin D_{HF} : h_{IA} + \frac{1}{2} \sum_{R \in D_{HF}} (g_{IARR} - g_{IRR}) = 0 \quad (24)$$

even though the individual integrals might have a high magnitude. The same $A \leftarrow I$ excitation from an excited determinant can have a large non-zero matrix element just because of the contraction over different two-electron integrals. This is at the core of why it is hard to approximate and precompute determinant-independent single-excitation sampling probabilities.

In the special case of localized MO bases, and under the assumption that $I$ and $A$ are from spatially distant orbitals, we have $|h_{IA}| \approx |g_{IARR}| \approx |g_{IRR}| \approx 0$, i.e., the single excitations become vanishingly small regardless of the determinant. In light of this property, we suggest the approximation of contracting over all orbitals ($R$) for the weight $S_i^A$

$$S_i^A = \begin{cases} |h_{AI}| + \sum_R |g_{AIRR} - g_{ARR}| & I \neq A \\ 0 & \text{else} \end{cases} \quad (25)$$

Due to the triangle inequality ($|A + B| \leq |A| + |B|$) it follows that for all determinants $|\hat{H}_{(D_i)}(A \leftarrow I)| \leq S_i^A$. Next, as for the doubles (see eq. 17), we contract over all orbitals, thus further approximating the exact expression for single excitations (eq. 23), to yield

$$p_{PCHB}^{A}(I) = \frac{\sum_C S_i^C}{\sum_{CL} S_L^C} \quad p_{PCHB}^{A}(A|I) = \frac{S_i^A}{\sum_C S_i^C} \quad (26)$$
In table 2 we list all possible combinations of probability distributions for single excitations. As in the case of doubles (see table 1), we can draw from constrained subsets and guarantee the sampled indices to be occupied or unoccupied. Sometimes it might be even beneficial to return back to uniform sampling if the probability distribution is uniform enough.

Table 2: The different sampling methods for single excitations and their scaling.

<table>
<thead>
<tr>
<th>Method</th>
<th>Distribution</th>
<th>Scaling</th>
</tr>
</thead>
<tbody>
<tr>
<td>fast-weighted</td>
<td>$p^{\text{PCHB}}_1(I), p^{\text{PCHB}}_1(A</td>
<td>I)$</td>
</tr>
<tr>
<td>fully-weighted</td>
<td>$p^{\text{PCHB}}_1(I)</td>
<td>_{I\in</td>
</tr>
<tr>
<td>uniform</td>
<td>$p^{\text{uni}}(I)</td>
<td>_{I\in</td>
</tr>
</tbody>
</table>

We have investigated whether equation 26 approximates well the exact probabilities, and which of the combinations from table 2 is the most efficient sampling scheme for single excitations. In figure 6 the $p^{\text{PCHB}}_1(a_{↑}\downarrow|1_{↑}\downarrow)|_{a_{↑}\notin|D_i\rangle}$ values are shown for the stack of ten benzene molecules with fragment localized MOs, (the probability of drawing the hole $a_{↑}$ after having drawn a $\uparrow$-particle from the first spatial orbital). For the exact probability (eq. 23) we assume that we start from the Hartree-Fock reference determinant, which is given by doubly occupying all $\pi$-orbitals and keeping all $\pi^*$-orbitals empty (blue line in figure 6) The orange line shows the corresponding approximated probability $p^{\text{PCHB}}_1(a_{↑}\downarrow|1_{↑}\downarrow)|_{a_{↑}\notin|D_i\rangle}$ (equation 26). Figure 6 shows good agreement between the approximated and the exact single excitations. It also confirms the exponential decay of matrix elements with the distance and the importance of excluding spatially distant orbitals, which would be unnecessarily sampled in an uniform scheme.

The probability for selecting the particle $p^{\text{PCHB}}_1(I)|_{I\in|D_i\rangle}$ has a relative deviation of 14% from the uniform distribution. As for the double excitation case, the particle, $I$, can safely be sampled uniformly, and

$$p_{|D_i\rangle}(A \leftarrow I) = p^{\text{uni}}(I)|_{I\in|D_i\rangle} \cdot p^{\text{PCHB}}_1(A|I)|_{A\notin|D_i\rangle}$$  (27)
Figure 6: The PCHB probability of the hole in a single excitation after having picked an \( \alpha \) electron in the first spatial orbital, \( p_1^{\text{PCHB}}(a_1|1_\uparrow) \), on a stack of ten benzene molecules. The blue line shows the exact probability (Eq. 23) assuming that the starting determinant \( |D_i\rangle = |2, 2, 2, 0, 0, 0; 2, \ldots \rangle \) is given by doubly occupying all \( \pi \)-orbitals and keeping all \( \pi^* \)-orbitals empty for each benzene molecule. The orange line shows the approximated PCHB probabilities for single excitations (Eq. 26). Only unoccupied \( \alpha \) spin-orbitals, i.e. allowed holes, are considered. The orbital indices are sorted by decreasing exact probability (blue line). (b) Logarithmic scale of (a). The horizontal dashed line gives the probability of the corresponding unconstrained uniform distribution \( p_1^{\text{uni}}(a_1|1_\uparrow) \) as reference. The value is obtained from the inverse of the number of empty \( \alpha \)-orbitals \( 1/(n_{\text{orb}}^\uparrow - N_e^\uparrow) = 1/30 = 3.3 \cdot 10^{-2} \).
is predicted to be the fastest sampling scheme for single excitations.

Apart from the choice of the combination of sampling schemes we would like to mention that also the singles benefit from the improved sampling from constrained distributions discussed in subsection 3.2.

5 Applications

Four test case applications are discussed in the following to numerically show the improved performance of the novel excitation generators: (a) a chain of 30 hydrogen atoms with atom-localized orbitals, therefore dominated by single excitations, demonstrates the capabilities of our new algorithm for sampling single excitations, (b) a stack of different numbers of benzene molecules with fragment-localized MOs serves as example for a system of fragments with highly localized correlation effects that is dominated by double excitations and highlights the improved sampling of double excitations, (c) an Fe-porphyrin model complex shows the efficiency of the new single excitation sampling for systems with mostly delocalized MOs, as long as there is some spatial separation, and (d) the N$_2$-dimer with Hartree-Fock orbitals shows how a uniform particle selection for the double excitations can improve the performance of PCHB.

5.1 Computational Details

Orbital optimization (e.a. Hartree-Fock method), their transformations (e.a. localization), integral evaluations, their transformation to the basis of active molecular orbitals and their storage on the formatted FCIDUMP$^{34}$ file were performed using OpenMolcas.$^{35,36}$ The new excitation generation has been implemented in a locally modified version of NECI,$^3$ and will be made publicly available via our open-source repository$^{37}$ upon publication of this work. The different sampling schemes are compared by their variance reduction per wall-clock time $(\bar{V} \cdot t)^{-1}$. FCIQMC dynamics were entirely run in stochastic mode, i.e. no semi-
stochastic propagation was enabled. For every system and every algorithm there were three
calculations of 24 h, the efficiency was then averaged over these three runs. Since the variance
of the projected energy has to be estimated at stationary conditions, all calculations started
from walker population distributions of already converged wave functions.\textsuperscript{38} To guarantee
the optimal $\Delta \tau$, $p_1$, and $p_2$ for each algorithm we allowed $5 \cdot 10^4$ iterations of continued
$\tau$-search for every calculation, only iterations afterwards were considered for the efficiency.
The shift was adapted via a second-order scheme\textsuperscript{39} to keep the population close to the
target population of $1 \cdot 10^5$ for every calculation. The initiator criterion was applied with
a population threshold of three.\textsuperscript{17–19} Although we used the initiator approximation in all
applications presented here, we would like to emphasize, that our new excitation generator
applies also to the original (non-initiator) FCIQMC. The data analysis was performed using
pyblock, pandas, numpy, and matplotlib.\textsuperscript{40–43} All efficiencies were normalized to the current
PCHB algorithm in NECI. Sampling from constrained subsets uses our improved redrawing
algorithm (subsection 3.2) everywhere; more details in the appendix A.1.

5.2 Efficiency Comparison across Excitation Generators

Table 3 concisely summaries the efficiency of different combinations of sampling schemes.
We refer to the following subsections for a detailed discussion on the results for each test
case.

Uniform, fast-weighted, and fully-weighted sampling schemes for particles and holes are
labeled as unif, fast and full, respectively. The colon symbol (:) separates particle- from
hole-selection. For example, the sampling scheme for a double excitation, $p_{2\text{ini}}^\text{unif}(I) | I \in |D_i\rangle \cdot
p_{2\text{PCHB}}^\text{unif}(J|I) | J \in |D_i\rangle \cdot p_{2\text{PCHB}}^\text{fast}(A|IJ) \cdot p_{2\text{PCHB}}^\text{fast}(B|IJA)$, would be written as unif-full:fast-fast.

In principle there are $3^6 = 729$ different combinations of samplers, but we restrict our-
selves to the handful in table 3, which represent the most meaningful. The conditional
probabilities for the holes and for the second particle can use more information as compared
to the first particle, hence the later particle and hole selections should use less uniform sam-

26
pling. It follows that a \texttt{full:unif} single excitation generator is expected to be less efficient than \texttt{unif:full}, due to the unnecessary weighting of the first particle, which as shown earlier in general exhibits uniform distributions.

Table 3: The efficiency of different combinations of sampling schemes from table 1 and table 2 for all systems. The efficiency is given by $(V \cdot t)^{-1}$ and normalized to the currently used algorithm in NECI, which is \texttt{unif:unif unif-unif:fast-fast}. The fastest speedup factor per system is printed bold. The algorithms are lexicographically sorted by sampling quality \texttt{unif < fast < full}. The error for the last digit is given in parentheses.

<table>
<thead>
<tr>
<th>Excitation generation</th>
<th>Efficiency ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$H_{30}$</td>
</tr>
<tr>
<td>\texttt{unif:unif}</td>
<td>\texttt{unif-unif:fast-fast}</td>
</tr>
<tr>
<td>\texttt{unif:unif}</td>
<td>\texttt{unif-full:fast-fast}</td>
</tr>
<tr>
<td>\texttt{unif:unif}</td>
<td>\texttt{full-full:fast-fast}</td>
</tr>
<tr>
<td>\texttt{unif:fast}</td>
<td>\texttt{unif-unif:fast-fast}</td>
</tr>
<tr>
<td>\texttt{unif:fast}</td>
<td>\texttt{full-full:fast-fast}</td>
</tr>
<tr>
<td>\texttt{unif:full}</td>
<td>\texttt{unif-full:fast-fast}</td>
</tr>
<tr>
<td>\texttt{unif:full}</td>
<td>\texttt{unif-full:full-full}</td>
</tr>
<tr>
<td>\texttt{full:full}</td>
<td>\texttt{full-full:fast-fast}</td>
</tr>
<tr>
<td>\texttt{full:full}</td>
<td>\texttt{full-full:full-full}</td>
</tr>
</tbody>
</table>

\begin{itemize}
  \item \texttt{unif:unif:fast-fast} \footnote{The choice in \textsc{NECI}$^3$.}
  \item \texttt{unif-full:fast-fast} \footnote{The choice by Holmes et al.$^1$.}
\end{itemize}

The $H_{30}$-chain. The first example is a neutral chain of 30 hydrogen atoms that are equally spaced $4.2 \alpha_0$ apart. The used basis was STO-3G, which contains a minimal basis set of 30 basis functions, implying that (30,30) is the full-CI space. A $C_1$ point-group was utilized. Closed shell Hartree-Fock optimized orbitals were transformed by a Pipek-Mezey localization\textsuperscript{44} to obtain orthonormal and localized MOs.

Because of the fully atom-localized MOs the system is dominated by single excitations. If for example all excitations are selected in a fully-weighted manner, then $p_1$ is adjusted to 98\% by equation 13, i.e. mostly single excitations are attempted. This means that the choice
of the double excitation generator does not have a large influence on the overall efficiency. In table 3 we indeed see that the efficiencies are grouped by the excitation generator for the single excitations. The purely uniform generation for single excitations \texttt{unif:unif} performs worst, since it samples far away holes $A$ with the same probability as close ones. The fully weighted excitation generation \texttt{full:full} performs best \emph{per iteration}, but it is outperformed by \texttt{unif:full} when taking wall-clock time into account, since the choice of the first particle is near-uniformly distributed (see section 4). The CAS(30,30) has no excess of empty orbitals and surprisingly \texttt{unif:fast} is the fastest combination. The fast-weighted hole-selection \texttt{unif:fast} performs better (by roughly $2.9^{2.2} \approx 30\%$) than the fully-weighted hole-selection \texttt{unif:full}.

This can be explained as follows: Due to hermiticity (compare equation 25) we have $S^A_I = S^I_A$. If an excitation $A \leftarrow I$ has a high weight, then the reverse excitation $I \leftarrow A$ has a high weight as well. If $I$ gets only occupied via the excitation from $A$, then $A$ is unoccupied when $I$ is occupied. Since the single excitations are in good approximation the only excitations, there is no other source that would spawn particles to $I$ except the few nearest neighbors apart from $A$. Hence the probability of $A$ being unoccupied when $I$ is occupied is not independent from the precomputed weights, but rather correlated $p(A \notin |D_i| |I \in |D_i|) \propto p^{\text{PCHB}}_{I}(A|I)$. Numerically we confirm that if we sample via $p^{\text{PCHB}}_{I}(A|I)$ the rate of valid excitations is around 75%. If we would draw the hole $A$ uniformly without guaranteeing the orbital to be unoccupied (i.e. $p^{\text{uni}}_{I}(A|I)$) the rate of valid excitations would be $\frac{14}{29} \approx 50\%$.

The precomputed weighted single excitations improve the performance by a factor of nearly 3 compared to \texttt{unif:unif}. Since the double excitations do not play a large role in this system, the NECI and Holmes et al. approaches, differing exclusively in the double excitations, exhibit a very similar efficiency.\textsuperscript{1,3}
Stack of benzene molecules. The next example are two stacks of benzene at a distance of 3.0 Å with five and ten molecules, respectively. The stack of ten molecules is displayed in Figure 7. In both cases the orbitals were fragment-localized, i.e. natural, Hückel-like orbitals on each benzene. The active space contained all the $\pi$ and $\pi^*$ orbitals, i.e. a CAS(30,30) for five benzene and a CAS(60,60) for ten benzene molecules. As we have shown in a previous publication already single excitations between the fragments suffice to recover Full CI energy, while the system as a whole is dominated by double excitations inside each fragment.\textsuperscript{10}

In the case of the five molecules we used the CASSCF(30,30) optimized orbitals from our previous work.\textsuperscript{10} The MOs for the stack of ten fragments were generated by doubling the coefficients along the diagonal to form a block-diagonal coefficient matrix of two blocks. Since the molecular orbitals of this block-diagonal matrix are not orthonormal, a Gram-Schmidt orthonormalisation was performed prior to the stochastic CASSCF optimization. This implies that the CAS(30,30) is at the energetic minimum with respect to orbital rotations, while the CAS(60,60) is only close to it. Hence the relevance of inter-fragment single excitations, recovering orbital rotations, is enhanced for the CAS(60,60). In the first case we have $p_2 = 78\%$, while in the second case $p_2 = 67\%$, when sampling everything fully-weighted.

In table 3 the efficiencies for different sampling schemes are listed for both stacks of benzene. As expected from the discussion in section 3 and 4 we see that uniform selection of the first particle and subsequent fully-weighted sampling, i.e. unif:full unif-full:full-full, works best for both types of systems. It is 3.7 faster than the current choice in NECI and 60 % faster than the choice by Holmes et al.. The advantage of not sampling uniformly increases with system size, in fact the number of direct neighbors stays the same while the number of
distant orbitals increases. We also see that, unlike the H$_{30}$ chain, it is important to improve the sampling of both single and double excitations. For example varying the sampling of doubles by going from unif:full full-full:fast-fast to unif:full unif-full:full-full improves the performance from 2.6 to 3.7.

For these active spaces with a similar number of particles and empty orbitals the fully-weighted hole-selection outperforms the fast-weighted hole-selection, but the difference is perhaps less pronounced than expected. Similar to the H$_{30}$-chain this can be explained as follows: Again, due to hermiticity, we have $W_{IJ}^{AB} = W_{AB}^{IJ}$ (see equation 15). If an excitation $AB \leftarrow IJ$ has a high weight, then the reverse excitation $IJ \leftarrow AB$ has a high weight as well. Hence the probability of $A$ and $B$ being unoccupied when $I$ and $J$ are occupied is not independent from the precomputed weights, but rather correlated, i.e. $p(A \notin |D_i\rangle |I, J \in |D_i\rangle) \propto p_{\text{PCHB}}^2(A|IJ)$. We can confirm this correlation numerically. The hypothetical, uncorrelated rate of a valid hole selection for this active space would be roughly $1/2 \cdot 1/2 = 1/4$, since half of the orbitals are valid picks for the first hole and (half - 1) of the orbitals would be valid picks for the second hole. The numerically obtained rate of valid excitations for fast-weighted hole selection is roughly 50%.

Contrary to the H$_{30}$ chain the fully-weighted hole-selection is still more efficient than the fast-weighted one, since there are other effects, that lower the aforementioned correlation between PCHB weights and occupation. There are both single and double excitations and there can be double excitations of the type $IJ \leftarrow AR$. Hence it is more probable that excitations happen to $I$ or $J$ that do not empty $A$ and $B$ simultaneously.

**Fe(II)-porphyrin.** The next system is a Fe(II)-porphyrin model complex whose properties were intensely discussed in our previous publications.$^{4,5,9,10,45,46}$ We use the converged stochastic CASSCF natural orbitals from our previous work$^9$ with a (32,34) active space that consists of nine doubly-occupied $\pi$-orbitals, seven empty $\pi^*$-orbitals, five metal centered 3$d$-orbitals, four doubly-occupied $\sigma_N$-orbitals, four empty orbitals of the (4$s$4$p$) shell,
and five empty $d'$-orbitals (double-shell orbitals). The natural MOs for this system are delocalized, but not over the whole molecule. They are still confined to certain regions and can be identified as metal-centered, aromatic, or $\sigma$-donating nitrogen orbitals. So there is spatial separation of MOs, but not as strongly as for the stack of five or ten benzene molecules.

In table 3 the efficiencies of different sampling schemes are listed for the Fe(II)-porphyrin molecule. As for the stack of benzene molecules the fastest choice is again **unif:full** unif-full:full-full, but the maximum speedup by a factor of 1.21 is much smaller than for the localized benzene stack.

N$_2$-dimer. The last example is Full CI on the N$_2$ dimer at equilibrium distance with a Dunning’s cc-pVQZ basis\textsuperscript{33} using Hartree-Fock orbitals, resulting in a (14,110) active space. This system is highly delocalized and has an excess of empty orbitals. The excess of empty orbitals means that a randomly picked index is likely unoccupied, e.g. $p^{PCHB}_2(A|IJ) \approx p^{PCHB}_2(A|IJ)_{A\notin|D_i}$, which means that the fast-weighted hole-selection outperforms the fully-weighted one. On the other hand it is very important to guarantee orbitals to be occupied in the particle-selection, for both single and double excitations. This means that only uniform or fully-weighted particle-selection are expected to be performant and explains why **unif-fast** particle selection for double excitations performs so badly for this system (table 3).

In the case of single excitations, the approximation of contracting over all orbitals when calculating the weights (equation 25) is particularly poor for these Hartree-Fock orbitals, which means that PCHB weighting for single excitations does not have the same benefits as for the other systems. For this reason we see that purely uniform sampling (**unif:unif**) and uniform sampling of particles followed by fast-weighted selection of holes (**unif:fast**) are performing very similarly.

In the case of double excitation there is no approximation regarding the matrix elements that would break down because of the use of Hartree-Fock orbitals. This means that the PCHB weights for doubles are close to the correct weights and from the discussion so far
we conclude that fast-weighted hole-selection is the fastest method, which we can confirm in table 3. The probabilities for the first particle $I$ are near-uniformly distributed, (compare discussion around figure 4), hence the fastest double particle-selection is either unif-unif or unif-full. The first one is faster per wall-clock time, the latter samples with higher quality. Looking at table 3 we see that purely uniform particle selection (unif-unif) is outperforming unif-full, which makes unif:unif unif-unif:fast-fast the fastest method for this system.

6 PCHB and GAS

In the following we show how, in the stochastic variant of the Generalized Active Space approach (Stochastic-GAS), occupation number constraints can be incorporated into the new weighted sampling scheme for singles and how the weighted particle-selection for single and double excitations can improve the GAS performance. In the following we offer a brief summary of the GAS concept. Details on the Stochastic-GAS algorithm can be found in Reference 10.

In GAS the active orbitals are partitioned into a number of active subspaces. Then only some distributions of particles among these subspaces are allowed, for example by defining a minimum and maximum particle number per GAS space. A given allowed distribution of particles among the subspaces is called a supergroup.\textsuperscript{10,13,47} For a given supergroup all possible configurations inside each subspace are allowed, i.e. a full CI wave function inside each subspace is generated. If there is only one supergroup the GAS spaces are disconnected and no excitations between them are allowed. In figure 8 a possible GAS wave function is displayed. The GAS algorithm is a natural generalization of other truncated FCI schemes and it can easily be used to construct restricted active spaces (RAS) (three subspaces) or complete active spaces (CAS) wave function expansions.\textsuperscript{13,48–51}

As we showed in our previous work\textsuperscript{10} the GAS constraints can be included into the
Figure 8: An example GASCI wave function.

PCHB probability distributions with negligible additional run time cost. The supergroup of a given determinant can be determined by counting the particles per GAS space (a $\mathcal{O}(N_e)$ operation). Counting how many particles an excitation transfers between GAS spaces is a trivial operation. Hence for a given supergroup (and all determinants belonging to it), an excitation is GAS allowed if the resulting particle distribution is still inside the chosen GAS constraints. This implies in particular that the starting determinant $|D_i\rangle$ is not necessary, only its supergroup is required to determine if an excitation is GAS allowed.

For every supergroup we can then define modified PCHB probability tables that automatically incorporate the GAS constraints. If we define $i_{sg}$ to be a labeling index for the supergroups, we can introduce modified PCHB weights, based on equations 25 and 15, that are parameterized by the supergroup

$$
\tilde{S}_{I}^{A}(i_{sg}) = \begin{cases} 
S_{I}^{A} & (A \leftarrow I) \text{ GAS allowed for } i_{sg} \\
0 & \text{else} 
\end{cases}
$$

$$
\tilde{W}_{IJ}^{AB}(i_{sg}) = \begin{cases} 
W_{IJ}^{A} & (AB \leftarrow IJ) \text{ GAS allowed for } i_{sg} \\
0 & \text{else} 
\end{cases}
$$

At run time we have to determine the supergroup index for a given determinant, the details
of how to do this efficiently are in our previous work,\textsuperscript{10} and then draw from the corresponding probability distributions that were precomputed from eq. 28.

Apart from the user defined GAS constraints there is, due to the Pauli-principle, a "natural" maximum particle number per GAS space, namely the number of spin-orbitals. The weighted particle selection based on the GAS PCHB weights (equation 28) can use that information and improve the sampling performance.

To see this, let us consider the Anderson impurity model which consists of one impurity orbital and \( n_{\text{bath}} \) bath orbitals.\textsuperscript{52} The model allows only single excitations and only between the impurity and the bath orbitals, i.e. there may not be any excitations between bath orbitals. If we start from a determinant, where the impurity is doubly occupied and select a particle on a bath orbital to excite from, there will be no valid single excitation possible. Since there is a uniform probability of \( \frac{N_e-2}{N_e} \) to select a particle from the bath orbitals, a uniform particle selection will produce invalid excitations with a probability of at least \( \frac{N_e-2}{N_e} \) for such determinants.

This inefficient sampling can be cured with GAS and weighted particle-selection. If we define the impurity orbital to be its own GAS space and the bath orbitals to be in a second GAS space, then there are three allowed supergroups \([2, N_e - 2], [1, N_e - 1], [0, N_e]\), where we just listed the number of particles per GAS space. If we start now from the supergroup \([2, N_e - 2]\) and assume that \( K \) is a spin-orbital index in the bath-orbitals, then we have \( \tilde{S}_K^A(i_{\text{sg}}) = 0 \) for all possible holes \( A \). This implies that \( p_{1}^{\text{PCHB}}(K|i_{\text{sg}}) = 0 \) and only one of the two particles in the impurity is selected to be excited; each one with a probability of \( \frac{1}{2} \).

## 7 PCHB and spin purification

In this section we want to show how the modified Hamiltonian

\[
\hat{H}' = \hat{H} + J \cdot \hat{S}^2, \quad 0 < J \in \mathbb{R}
\]

(29)
is automatically efficiently sampled via PCHB. We showed in a previous work how this first-order spin penalty is useful, particularly in FCIQMC, to calculate spin-pure states in a Slater-determinant basis.\textsuperscript{11}

Similarly to the matrix elements of the Hamiltonian (equations 5, 6, and 7) we write the matrix elements for the $\hat{S}^2$ operator as

$$
\begin{align*}
\hat{S}^2_{[D_i]}(\text{diag}) &= \langle D_i | \hat{S}^2 | D_i \rangle = \hat{S}_z(\hat{S}_z - 1) + n^{OS}_\alpha \\
\hat{S}^2_{[D_i]}(A \leftarrow I) &= \langle D_i | \hat{S}^2 | a_A^\dagger a_I D_i \rangle = 0 \\
\hat{S}^2_{[D_i]}(AB \leftarrow IJ) &= \langle D_i | \hat{H} | a_A^\dagger a_B^\dagger a_J a_I D_i \rangle = \begin{cases} \\
\Gamma_{AI}^{D_i} \Gamma_{BJ}^{D_{ij}} & \text{exchange excitation} \\
0 & \text{else}
\end{cases}
\end{align*}
$$

where $n^{OS}_\alpha$ is the number of open-shell $\alpha$ electrons of a given determinant and exchange excitations are all excitations of the type $i_{\alpha j\beta} \leftarrow i_{\beta j\alpha}$. The diagonal term does not appear in the excitation generation and the single excitations are not modified at all. Only the double excitations of exchange type are actually modified and their respective PCHB weight (equation 15) becomes $|g_{AIBJ} - g_{AJBI} + J|$. If we additionally assume spatial orbitals, i.e. $g_{AIBJ} = g_{a,i,b,j} \delta^{\sigma\tau} \delta^{\mu\nu}$, we finally get $|J - g_{ijji}|$.

For a given spin-projection $\sigma$ we write the primed $\sigma'$ for the opposite spin-projection. We conclude that the particle selection $P^\text{PCHB}_2(j_\sigma | i_{\sigma'})$ with the modified weight becomes proportional to the spin penalty, i.e. opposite spin pairs are more likely to be picked. The increased probability of exchange interactions depends continuously on $J$ and still incorporates the value of the integrals, e.g. $g_{ijji}$. The same applies to the subsequent hole-selections which are more likely to sample exchange excitations. Hence the modified hamiltonian is efficiently sampled, with a continuous dependence on $J$. 

35
8 Conclusions

In this work we have introduced an improved PCHB algorithm in FCIQMC that takes advantage of the locality of electron correlation, and that allows FCIQMC dynamics that are 2 to 4 times more efficient than currently available algorithms.

A weighted sampling method for single excitations has been presented that uses precalculated probabilities. The weights for single excitations are obtained by contracting over all orbitals in the two-electron term, rather than over only the occupied ones. This approximation works particularly well for localized orbitals, where it can take advantage of the exponential decay of integrals with increasing spatial distance. Precisely in the localized basis single excitations are pivotal for efficient FCIQMC dynamics, hence right when weighted sampling of single excitations becomes relevant our approximation works well.

Also, optimal combinations of uniform, fully-weighted, or fast-weighted sampling of particles or holes in double excitations have been identified for a variety of chemical systems, active spaces, and forms of the orbital basis (localized or delocalized). We conclude the following qualitative rules: (a) For localized orbitals and active spaces featuring a similar number of particles and spatial orbitals the \texttt{unif:full unif-full:full-full} is generally the fastest combination of sampling schemes, as shown by the fragment-localized benzene stacks and the Fe-porphyrin examples. The speedup was up to a factor of four compared to the current implementation in \texttt{NECI}. (b) If the system uses delocalized orbitals with an excess of empty ones, such as the N$_2$ dimer in a large basis, then \texttt{unif:unif unif-unif:fast-fast} is preferred. This combination gives a 50\% speedup compared to the choice of Holmes et al..

We have also suggested an alternative method to draw from constrained subsets of precalculated probabilities that is 5 to 10 times faster than the sampling methods currently available, and scales practically constant, when keeping the particle number constant and increasing the basis set size. The details are discussed in the appendix. This improvement can be used in all implementations of PCHB, for both singles and doubles and improves the performance of fully-weighted sampling.
Our new precomputed weighting scheme for single excitations can already be applied in combination with the Stochastic-GAS and the spin purification approach in FCIQMC, making both methods more efficient.

As a final remark, we would like to highlight that the full strength of the novel precomputed weighting scheme emerges when combined with the spin-adapted implementation of FCIQMC, based on the graphical unitary group approach (GUGA). This will be the subject of a future publication. In a series of recent works we have shown that by simple orbital localizations and reorderings it is possible to obtain CI Hamiltonian matrices with a unique and unprecedented block diagonal structure. As a consequence extremely compact wave functions (to the limit of near-single-reference) are obtained for ground and excited states of exchange-coupled poly-nuclear transition metal clusters. Since localized bases are propaedeutic to the GUGA wave function compression, it is clear that an excitation generator that exploits locality would further facilitate the screening of unnecessary excitation attempts in such sparse Hamiltonians. Our non-uniform drawing scheme in GUGA would promote those single and double excitations among nearest-neighbors and would screen out excitations that are related to decaying integral values, yielding efficiently sampled excitations in extremely sparse Hamiltonians.
A Appendix

A.1 Constrained sampling

In this section we will discuss the various ways of how to efficiently sample from constrained subsets of known discrete probability distributions. To do so we need to recapitulate the two main algorithms for sampling non-uniform, discrete distributions in the FCIQMC community, namely the cumulative distribution function (CDF) and the alias sampling\textsuperscript{23,32} methods.

The CDF sampling is shown in algorithm 1. It has linear scaling for the initialization and logarithmic scaling for the sampling step.

\textbf{Algorithm 1:} Cumulative Distribution Function sampling

\textbf{Data:} $w_i \geq 0$ unnormalized weights of length $n$. We assume zero-based indexing.

\textbf{Result:} A randomly returned index with its probability $p_i = \frac{w_i}{\sum_j w_j}$.

1.1 \textbf{Initialization}

1.2 $N_{\text{norm}} \leftarrow \sum_i w_i$; \hspace{1cm} // Scales as $O(n)$

1.3 \textbf{for} $0 \leq i < n$ \textbf{do}

1.4 $p_i \leftarrow \frac{w_i}{N_{\text{norm}}}$; \hspace{1cm} // Scales as $O(n)$

1.5 $p^\text{cum} \leftarrow \text{cumulative\_sum}(p)$; \hspace{1cm} // Scales as $O(n)$

1.6 \textbf{Sampling}

1.7 $r \leftarrow \text{get\_random\_number}([0,1])$;

1.8 $i \leftarrow \text{first\_index\_greater\_equal}(r, p^\text{cum})$; \hspace{1cm} // Scales as $O(\log(n))$

1.9 \textbf{return} $(i, p_i)$;

The alias sampling is shown in algorithm 2. The (cumulative) summation over weights $w_i$ is the same for both methods, but there is additional work in the initialization step of the alias sampling. This means that the scaling in the initialization is at least linear for both methods, but CDF sampling has a smaller prefactor. In the final step of the alias sampling (algorithm 2) there are two if-else-branches. Both branches have constant scaling, but the one that returns $K_{[r]}$ has an additional indirection. To avoid this case the sum of $U_i$ has to be maximized in the initialization, which is an NP-hard problem, but can be approximately achieved by pairing large $U_i$ with small $U_i$.\textsuperscript{54} Adding this optional sorting step lets the initialization of the alias tables scale with $O(n \log(n))$, but the sampling becomes
faster.

In addition there is another trade off between fast construction and fast sampling for both methods. Often the calling code, in particular FCIQMC, does not want to only sample a random value, but to also know its probability \( p_i \). The necessary renormalization of the weights \( p_i = \frac{w_i}{\sum_j w_j} \) does not have to be done at initialization, it could be computed upon request of the calling code.

To illustrate these effects we will first compare two implementations of the alias method. The STL \texttt{std::discrete\_distribution} implementation of GCC 7.5.0 is optimized for construction, while our own implementation is optimized for sampling time. The C++ code that was used to generate the data is available in the supporting information. The code ran on one core of an AMD EPYC 7763 CPU with 2.0 GHz.

In Figure 9 we see that there is a trade off between optimizing for construction or sampling time. For this reason we now assume that we optimize for sampling when the sampler is initialized once and then sampled often, while we optimize for construction, if the alias sampler should be constructed often.

Now we can turn our eyes towards an efficient algorithm for sampling from a constrained subset of an already constructed alias sampler. The main idea (algorithm 3) is to just sample via existing alias tables until the sampled index is contained in the subset. Since the alias sampling is so fast, this is usually faster than rebuilding alias or CDF tables for the constrained subsets.

There are some crucial details to improve the scaling, to correctly deal with floating point arithmetic, and to correctly work with zero probability distributions. If we are assuming that the index \( i \) is contained in the constrained subset \( M \), it is drawn with a probability \( \tilde{p}_i = \frac{p_i}{\sum_{j \in M} p_j} \) from it. The renormalization scales linearly with the number of elements in \( M \). If \( M \) contains many elements it is advantageous to scale with the complement via

\[
\sum_{j \in M} p_j = 1 - \sum_{j \notin M} p_j .
\]
There are certain situations when all weights and all $p_i$ are zero, then the application of equation 31 would wrongly yield $\sum_{j \in M} p_j = 1$ although it should be zero. This has to be distinguished from the case where it is indeed $\sum_{j \in M} p_j = 1$. For this reason we have to directly evaluate $\sum_{j \in M} p_j$, if the right hand side of equation 31 is exactly one. There is also another reason that necessitates the direct evaluation. Equation 31 is not exactly true for floating numbers. Since the left-hand-side appears in the denominator for renormalization (compare algorithm 3), these small differences can “blow up”, if the overall expression is small. For this reason the direct calculation is preferred, if $N_{\text{renorm}} \approx 0$. Both cases are caught in line 3.6.

If $N_{\text{renorm}}$ is nonzero but small, then algorithm 3 will perform many iterations in the while-loop (line 3.14). In this case it is advised to switch back to CDF sampling (See line 3.10).

It is advantageous to represent $M$ as both a sorted enumeration of the contained indices and as a bit mask, since both have an advantage for certain operations. The bit mask allows a very fast lookup to determine $i \in M$ and the complement is just the bit wise NOT. The sorted enumeration can be faster to perform restricted summations over an index.

In figure 10 we show the performance of constrained sampling via redrawing (algorithm 3) and compare against sampling via rebuilding the CDF or alias tables. The rebuilt alias tables are optimized for initialization, i.e. there is no sorting step in algorithm 2. We restrict the system size to 500, since we assume that calculations in the near future will not exceed 500 spin orbitals. The subset size, measured as fraction of the number of elements, is varied with the following values $\{0.25, 0.5, 0.9\}$. As theoretically expected, all methods scale linearly with the system size but there is considerable difference between redrawing and rebuilding. Redrawing is near-independent from the subset fraction, since the resampling in the while-loop of algorithm 3 is so fast. Only if the subset weight gets really small it has to switch to a rebuilding algorithm (see line 3.10). The rebuilding algorithms on the other hand depend on the subset fraction; the higher the subset fraction the slower they become. The difference
between CDF and alias sampling is negligible for rebuilding and drawing once. If the subset fraction is 0.5, which is the usual case for balanced active spaces with equal number of spatial orbitals and particles, then redrawing is 5 to 10 times faster than rebuilding.

Figure 9: Sampling times of the alias method with- and without the optional sorting step in the initialization (compare algorithm 2) against the system size $n$. The dotted lines are purely the sampling times while the solid lines include the sampling and construction time.
Algorithm 2: Alias sampling

Data: $w_i \geq 0$ unnormalized weights of length $n$. We assume zero-indexation.

Result: A randomly returned index with its probability $p_i = \frac{w_i}{\sum_j w_j}$.

2.1 Initialization

2.2 $N_{\text{norm}} \leftarrow \sum_i w_i$; \hspace{1cm} // Scales as $\mathcal{O}(n)$
2.3 overfull $\leftarrow \emptyset$; \hspace{1cm} \hspace{1cm} // Initialize as empty vector
2.4 underfull $\leftarrow \emptyset$;
2.5 \hspace{1cm} for $0 \leq i < n$ do \hspace{1cm} // Scales as $\mathcal{O}(n)$
2.6 \hspace{2cm} $K_i \leftarrow i$;
2.7 \hspace{2cm} $p_i \leftarrow \frac{w_i}{N_{\text{norm}}}$;
2.8 \hspace{2cm} $U_i \leftarrow n \cdot p_i$;
2.9 \hspace{2cm} if $U_i < 1$ then
2.10 \hspace{3cm} append(underfull, $i$);
2.11 \hspace{2cm} else
2.12 \hspace{3cm} append(overfull, $i$);
2.13 \hspace{1cm} Optional sorting of index by value of $U_i$
2.14 \hspace{2cm} underfull $\leftarrow \text{sort}(\text{underfull}, \text{key}=i \mapsto U_i)$; \hspace{1cm} // Scales as $\mathcal{O}(n \log(n))$
2.15 \hspace{2cm} overfull $\leftarrow \text{sort}(\text{overfull}, \text{key}=i \mapsto U_i)$; \hspace{1cm} // Scales as $\mathcal{O}(n \log(n))$
2.16 \hspace{1cm} while (size(underfull) $\neq 0$) $\land$ (size(overfull) $\neq 0$) do \hspace{1cm} // Scales as $\mathcal{O}(n)$
2.17 \hspace{2cm} // Pairs very small with very large, if sorted.
2.18 \hspace{2cm} $s \leftarrow \text{pop\_first}(\text{underfull})$;
2.19 \hspace{2cm} $l \leftarrow \text{pop\_last}(\text{overfull})$;
2.20 \hspace{2cm} $K_s \leftarrow l$;
2.21 \hspace{2cm} $U_l \leftarrow U_l - (1 - U_s)$;
2.22 \hspace{2cm} if $U_l < 1$ then
2.23 \hspace{3cm} prepend(underfull, $l$);
2.24 \hspace{2cm} else
2.25 \hspace{3cm} prepend(overfull, $l$);
2.26 \hspace{2cm} while size(underfull) $\neq 0$ do
2.27 \hspace{3cm} $s \leftarrow \text{pop\_last}(\text{underfull})$;
2.28 \hspace{3cm} $U_s \leftarrow 1$;
2.29 \hspace{2cm} while size(overfull) $\neq 0$ do
2.30 \hspace{3cm} $l \leftarrow \text{pop\_last}(\text{overfull})$;
2.31 \hspace{3cm} $U_l \leftarrow 1$;
2.32 Sampling
2.33 $r \leftarrow \text{get\_random\_number}([0, n])$;
2.34 \hspace{1cm} if $(r - \lfloor r \rfloor) < U_{\lfloor r \rfloor}$ then
2.35 \hspace{2cm} $i \leftarrow \lfloor r \rfloor$;
2.36 \hspace{1cm} else
2.37 \hspace{2cm} $i \leftarrow K_{\lfloor r \rfloor}$;
2.38 \hspace{1cm} return $(i, p_i)$;
Algorithm 3: Constrained sampling via redrawing.

Data: An initialized alias sampler object $S$ and a constrained subset $M$ of the index.

Result: A randomly returned value $i$ from $M$ with its renormalized probability

$$\tilde{p}_i = \frac{p_i}{\sum_{j \in M} p_j}.$$

3.1 Initialization

3.2 if $\text{size}(M) < \frac{\text{size}(\text{index})}{2}$ then
3.3 $N_{\text{renorm}} \leftarrow \sum_{j \in M} p_j$; // Scales as $O(\text{size}(M))$
3.4 else
3.5 $N_{\text{renorm}} \leftarrow 1 - \sum_{j \notin M} p_j$; // Scales as $O(\text{size}(\text{index}) - \text{size}(M))$
3.6 if $N_{\text{renorm}} = 1 \lor N_{\text{renorm}} \approx 0$ then
3.7 $N_{\text{renorm}} \leftarrow \sum_{j \in M} p_j$; // Scales as $O(\text{size}(M))$
3.8 if $N_{\text{renorm}} = 0$ then
3.9 return $(\text{NaN}, \text{NaN})$;
3.10 if $N_{\text{renorm}} < T$ then // A good value for $T$ is $10^{-1}$.
3.11 $(i, p_i) \leftarrow \text{CDF} \_\text{sample}(p_i \in M)$;
3.12 return $(i, \frac{p_i}{N_{\text{renorm}}})$;
3.13 $(i, p_i) \leftarrow \text{sample}(S)$;
3.14 while $i \notin M$ do
3.15 $(i, p_i) \leftarrow \text{sample}(S)$;
3.16 return $(i, \frac{p_i}{N_{\text{renorm}}})$;

43
Figure 10: Sampling times for drawing from a constrained subsets. The colors denote the algorithm: blue uses redrawing (algorithm 3), orange and green rebuild an alias and CDF sampler for the constrained subset. The line style denotes the size of the subset measured as fraction of the number of elements.

Acknowledgement

The authors thank Robert Anderson, Philip Haupt, and Werner Dobrautz for valuable scientific discussions (all MPI FKF). OW thanks Kai Guthier for the clean code he left, which was easy to extend. The authors thank Armin Burkhardt, Armin Schuhmacher, and Jörg Tatchen (MPI FKF) for keeping our computer clusters up and running. All authors acknowledge the Max Planck Society for financial support.

Supporting Information Available

The supporting information is made available free of charge, containing the following items: (a) geometries, starting orbitals, OpenMolcas inputs, and NECI inputs for our application test cases, and (b) the code to time different (constrained) sampling strategies (data for figures 9 and 10).
References


(30) Ref. 25, p.399.

(31) Ref. 25, p.401.


(36) Li Manni, G. et al. The OpenMolcas Web: A Community-Driven Approach to Advancing Computational Chemistry. 2023; 10.26434/chemrxiv-2023-b7f0j-v2.


Graphical TOC Entry