VARIATIONAL MONTE CARLO BASICS AND APPLICATIONS TO ATOMS AND MOLECULES

C. J. UMRIGAR
Cornell Theory Center and LASSP,
Cornell University, Ithaca, NY 14853.

Abstract. The two ingredients of variational Monte Carlo, *viz.*, the Metropolis method for sampling any known probability distribution and the variance minimization method for optimizing trial wavefunctions are discussed. General issues, related to the efficiency of the methods, such as the choice of the Metropolis proposal matrix, are emphasised. The ideas are illustrated by examples from electronic structure calculations of atoms and molecules.

1. Introduction

The variational Monte Carlo method was first applied (to the best of my knowledge) by McMillan [1] to study bosonic systems (liquid ^{4}He) and then by Ceperley, Chester and Kalos [2] to fermionic systems. Although it is the simplest of the quantum Monte Carlo methods, significant choices have to be made in practical implementations of the method that can have a large effect on both the accuracy of the results obtained and the efficiency with which they are obtained. Accordingly, such issues will be emphasised in this chapter.

Variational Monte Carlo is usually used to calculate expectation values of operators $\hat{X}$ that are diagonal or near-diagonal in the $\mathbf{R}$ representation. (The case of an arbitrary operator $\hat{X}$ is treated in the chapter by Nightingale.) The expectation value can be written in the form

$$\begin{aligned} \langle \hat{X} \rangle_\rho &= \frac{\sum_{\mathbf{R}} X(\mathbf{R}) \rho(\mathbf{R})}{\sum_{\mathbf{R}} \rho(\mathbf{R})} \\ &\approx \frac{1}{T} \sum_{i=1}^{T} X(\mathbf{R}_i) \end{aligned} \tag{1}$$

where $\rho(\mathbf{R}) = |\Psi(\mathbf{R})|^2$ and $\Psi(\mathbf{R})$ is a trial state that approximates the eigenstate of interest. The *configurational eigenvalue* $X(\mathbf{R})$ is defined by

$$X(\mathbf{R})\Psi(\mathbf{R}) \equiv \langle \mathbf{R}|\hat{X}|\Psi(\mathbf{R})\rangle = \sum_{\mathbf{R}'}\langle \mathbf{R}|\hat{X}|\mathbf{R}'\rangle\langle \mathbf{R}'|\Psi\rangle, \tag{2}$$

and $\Psi(\mathbf{R}) \equiv \langle \mathbf{R}|\Psi\rangle = \langle \Psi|\mathbf{R}\rangle$. To be of practical use it is necessary that the sum on the right hand side of Eq. (2) can be performed quickly. If the states $\mathbf{R}$ form a continuum, then this is the case if $\hat{X}$ is diagonal or near-diagonal (involves only low-order derivatives in the $\mathbf{R}$ representation). The operator of greatest interest is the Hamiltonian $\mathcal{H}$, in which case the configurational eigenvalue $X(\mathbf{R})$ is the *local energy.*

The second line in Eq. (1) expresses the fact that a Monte Carlo estimation of expectation values is obtained as a time average over a sequence of T configurations that are sampled with relative probability $\rho(\mathbf{R})$. The time average is an approximation to the quantum mechanical expectation value, which becomes an equality in the $T \to \infty$ limit. To evaluate the right hand side of Eq. (1), we need to find a way to sample $\rho(\mathbf{R})$. The *generalized Metropolis* method, described in Section 2, is almost universally used to sample complicated probability distributions. Although it is a very simple method, we will see that a little bit of thought invested in the implementation of the method can have a large effect on its efficiency by reducing sequential correlations of the sampled configurations.

As the trial state $\Psi(\mathbf{R})$ approaches an exact eigenstate of $\hat{X}$, the configurational eigenvalue becomes a constant, equal to the eigenvalue of $\hat{X}$ for that state, *i.e.*, both the systematic and the statistical errors of the Monte Carlo estimate of Eq. (1) vanish. Although this ideal situation is never achieved in practice for problems of interest, it does indicate that significant gains can be achieved by optimizing $\Psi(\mathbf{R})$. In Section 3 we present the *variance minimization method* for optimizing trial states.

It is worth pointing out that any Monte Carlo method should be viewed as a method of last resort. Monte Carlo methods are never efficient in the sense that they have statistical errors that go as

$$\text{error} \sim \frac{C}{\sqrt{N}} \tag{3}$$

where N is the number of Monte Carlo moves and therefore a measure of the computer time used. It is often possible (say by employing a more efficient variant of the Metropolis method or by optimizing the trial states) to improve the proportionality constant C by orders of magnitude but the $1/\sqrt{N}$ behavior is usually [3] inescapable. This should be contrasted to deterministic computational methods which often converge as $1/N^p$, where

p is typically a small integer, or even as $e^{-\alpha N}$. For example, Simpson rule integration with N grid points converges asymptotically as $N^{-4/d}$ for sufficiently smooth integrands. For one-dimensional problems, the N^{-4} convergence is vastly more efficient than that of Monte Carlo methods but for $d > 8$ Monte Carlo integration is more efficient. Hence it should come as no surprise that in practice Monte Carlo methods are used mostly for problems in a large number of dimensions, where it is not feasible to use deterministic numerical quadrature. All other methods, that attempt to treat many-dimensional problems, make approximations that either reduce the problem to a collection of problems in a smaller number of dimensions or that include the effect of coupling terms only perturbatively.

2. Generalized Metropolis Method

The Metropolis method [4] was originally used to sample states or configurations from the thermal Boltzmann distribution and this is still one of its more common uses. However, it is in fact a general method for sampling any *known* probability distribution [5]. Many simple (usually few-dimensional) distributions can be sampled directly (see *e.g.* Refs. [6, 7]). Direct sampling methods, where feasible, are preferable to the Metropolis method, since the points generated by the latter are serially correlated, sometimes with very long correlation times. However, most many-dimensional distributions of interest cannot be sampled directly and the Metropolis method comes in very handy. It is one of the most frequently used methods in both classical and quantum Monte Carlo and is a component of some of the more sophisticated forms of quantum Monte Carlo, *e.g.*, path integral Monte Carlo.

The Metropolis method is an application of a *Markov chain*. A Markov chain is specified by two ingredients: an initial probability distribution and a transition matrix whose elements $M(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ represent the probability of making a transition from an initial state $\mathbf{R}_\mathrm{i}$ to a final state $\mathbf{R}_\mathrm{f}$. Clearly, since the probability must be non-negative and since the system must evolve to some state (including the initial state), we require that

$$M(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) \geq 0 \quad \text{and} \quad \sum_{\mathbf{R}_\mathrm{f}} M(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = 1. \tag{4}$$

Such a matrix will be referred to as a *stochastic matrix*.

To sample the desired distribution $\rho(\mathbf{R})$ we start from an arbitrary initial state $\mathbf{R}_\mathrm{i}$ and evolve the system by repeated application of a Markov matrix M that satisfies the *stationarity condition*:

$$\sum_i M(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\, \rho(\mathbf{R}_\mathrm{i}) = \sum_i M(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})\, \rho(\mathbf{R}_\mathrm{f}) = \rho(\mathbf{R}_\mathrm{f}) \quad \text{for all states } \mathbf{R}_\mathrm{f}, \tag{5}$$

where the second equality follows from Eq. (4). Hence, ρ is a right eigenvector of M with eigenvalue 1. The stationarity property implies that if we start with the desired distribution ρ, we will continue to sample the same distribution. However, we want that any initial distribution should evolve to the desired distribution ρ under the repeated application of M, *i.e.*, $\lim_{n\to\infty} M^n(\mathbf{R}'|\mathbf{R}) = \rho(\mathbf{R}')$, independent of $\mathbf{R}$. In other words, not only should ρ be a right eigenvector of M but it should be the *dominant* right eigenvector. Necessary conditions are that transitions can be made with non-zero probability between (almost [8]) any pair of states in a finite number of steps and that M not be cyclic [9]. We will refer to such a matrix as being *ergodic*. In practice, in order to obtain reliable results, it is important that the length of the Monte Carlo run be long enough that there be a significant probability of the system making several transitions between the neighborhoods of any pair of states, each of which makes a significant contribution to the average. This ensures that states are visited with the correct probability with only small statistical fluctuations.

The rate at which the initial distribution evolves to the desired distribution ρ and the autocorrelation time of estimates of various observables is governed by the subdominant eigenvalues. The ideal situation (realised in direct sampling methods, which as discussed before can be used for only a few simple distributions) is that all the other eigenvalues are zero. In that case a distribution of initial configurations evolves to ρ in a single Monte Carlo step and every measurement is independent. In practice, Markov matrices with short autocorrelation times for the variables of interest are constructed not by analyzing the eigenvalue spectrum of the matrix but by using heuristic arguments to devise matrices that allow the system to evolve rapidly through configuration space.

Note that both the Metropolis method and the projector Monte Carlo methods discussed in other chapters in this volume are stochastic implementations of the power method for projecting out the dominant eigenvector of a matrix. In the Metropolis method one designs a matrix M which projects out a known distribution, whereas in the projector Monte Carlo methods one projects out the (unknown) dominant eigenstate of a matrix that is known, though possibly only in some limit. In the next two subsections we describe how the freedom in the choice of M in the Metropolis method can be exploited.

2.1. DETAILED BALANCE CONDITION

We have shown so far that we can sample ρ by repeated application of a Markov matrix that satisfies the stationarity condition, Eq. (5), but we have not shown how such matrices are constructed in practice. In order to do this, the more stringent *detailed balance* condition

$$M(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\ \rho(\mathbf{R}_\mathrm{i}) = M(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})\ \rho(\mathbf{R}_\mathrm{f}) \tag{6}$$

is usually imposed. It expresses the condition that for any pair of states $\mathbf{R}_\mathrm{i}$ and $\mathbf{R}_\mathrm{f}$ the rate of flow is the same in both directions. It is a sufficient (but not necessary) condition. Clearly the detailed balance equation Eq. (6) implies the stationarity condition, Eq. (5), but the reverse is not true.

In order to proceed further, it is useful to write the elements of the Markov matrix M as the product of the elements of a proposal (attempted transition) matrix T and an acceptance matrix A,

$$M(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = A(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\ T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}). \tag{7}$$

Both $M(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ and $T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ are stochastic matrices. The detailed balance equation is now

$$A(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\ T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\ \rho(\mathbf{R}_\mathrm{i}) = A(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})\ T(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})\ \rho(\mathbf{R}_\mathrm{f}) \tag{8}$$

or

$$\frac{A(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})}{A(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})} = \frac{T(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})\ \rho(\mathbf{R}_\mathrm{f})}{T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\ \rho(\mathbf{R}_\mathrm{i})}\ . \tag{9}$$

2.2. CHOICE OF ACCEPTANCE MATRIX

There are an infinity of choices for the acceptance matrix that satisfy Eq. (9). In fact, any function

$$F\left(\frac{T(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})\ \rho(\mathbf{R}_\mathrm{f})}{T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\ \rho(\mathbf{R}_\mathrm{i})}\right) = A(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) \tag{10}$$

for which $F(x)/F(1/x) = x$ will do. The choice that maximizes the acceptance is the one prescribed by Metropolis *et al.*[4]

$$A(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = \min\left\{1, \frac{T(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})\ \rho(\mathbf{R}_\mathrm{f})}{T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\ \rho(\mathbf{R}_\mathrm{i})}\right\}. \tag{11}$$

In later work, other choices for $A(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ have sometimes been made, *e.g.*

$$A(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = \frac{T(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})\ \rho(\mathbf{R}_\mathrm{f})}{T(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})\ \rho(\mathbf{R}_\mathrm{f}) + T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\ \rho(\mathbf{R}_\mathrm{i})}. \tag{12}$$

but these lead to a slower evolution of the system. Note, however, that there exist certain lattice models for which the $A(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ of Eq. (11) cannot be used with sequential sampling [see item (4a) in subsection (2.4)] because it yields a non-ergodic $M(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$, but the $A(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ of Eq. (12) does not suffer from this problem.

2.3. CHOICE OF PROPOSAL MATRIX

Next we discuss what is an efficient choice for the proposal matrix T. This is where some thought can result in a large payoff in increased efficiency. We will present ideas developed in Ref. [10], but in greater detail. The original Metropolis method [4] employed a proposal matrix $T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ which is symmetric, in which case the factors of T in Eq. (11) cancel. Hence, we have called the method presented here, the *generalized* Metropolis method. Surprisingly, the first observation (to my knowledge) in the literature that $T(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})$ need not be symmetric did not occur until 1970 [11].

We write $T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ in a form that is manifestly normalized to unity,

$$T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = \frac{S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})}{I(\mathbf{R}_\mathrm{i})}, \quad \text{where} \quad I(\mathbf{R}_\mathrm{i}) = \int d\mathbf{R}_\mathrm{f}\, S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}). \tag{13}$$

Then the ratio of acceptance probabilities in Eq. (9) becomes

$$\frac{A(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})}{A(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})} = \frac{T(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})}{T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})}\frac{\rho(\mathbf{R}_\mathrm{f})}{\rho(\mathbf{R}_\mathrm{i})} = \frac{I(\mathbf{R}_\mathrm{i})}{I(\mathbf{R}_\mathrm{f})}\frac{S(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})}{S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})}\frac{\rho(\mathbf{R}_\mathrm{f})}{\rho(\mathbf{R}_\mathrm{i})} \tag{14}$$

and the acceptance probability in Eq. (11) becomes

$$A(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = \min\left\{\frac{I(\mathbf{R}_\mathrm{i})}{I(\mathbf{R}_\mathrm{f})}\frac{S(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})}{S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})}\frac{\rho(\mathbf{R}_\mathrm{f})}{\rho(\mathbf{R}_\mathrm{i})}, 1\right\}. \tag{15}$$

The points sampled by the Metropolis method are sequentially correlated, resulting in a loss of computational efficiency. The effective number of independent observations in a Monte Carlo run of length N is only N/T_{corr} where T_{corr} is the autocorrelation time of the observable of interest for the particular form of the Metropolis method used. It is clearly advantageous to reduce T_{corr}. This can be done by either increasing the average size of the proposed moves or by increasing the acceptance of the moves. In order to prevent the acceptance from getting too small, it is common practice to restrict the moves to be in the neighborhood of $\mathbf{R}_\mathrm{i}$ by choosing $S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ to be non-zero only within a domain $D(\mathbf{R}_\mathrm{i})$ of volume $\Omega(\mathbf{R}_\mathrm{i})$ around $\mathbf{R}_\mathrm{i}$. For a *given* functional form of $S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ the acceptance decreases as $\Omega(\mathbf{R}_\mathrm{i})$ (and therefore the average size of the proposed moves) increases, so, there exists an optimal $\Omega(\mathbf{R}_\mathrm{i})$ for which the system evolves the fastest.

Our task is to find a functional form for $S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ such that $I(\mathbf{R}_\mathrm{i})$ is known, $T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ can be sampled directly, the resulting proposed moves are large and the acceptance in Eq. (15) is large on average. If $I(\mathbf{R}_\mathrm{f}) = I(\mathbf{R}_\mathrm{i})$ for all states $\mathbf{R}_\mathrm{f}$ that are accessible from $\mathbf{R}_\mathrm{i}$ in a single step then the factors of $I(\mathbf{R}_\mathrm{i})$ and $I(\mathbf{R}_\mathrm{f})$ in Eq. (15) cancel and the choice $S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) \overset{\propto}{\sim} \rho(\mathbf{R}_\mathrm{f})$ results in $A(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})/A(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f}) \approx 1$. (The symbol $\overset{\propto}{\sim}$ should be read "approximately proportional to".) This is in fact the case in the *heat bath* algorithm, frequently used for discrete systems, discussed in Section 2.4. For most $\rho(\mathbf{R})$ of interest, it is not possible to find a function that approximates $\rho(\mathbf{R})$ sufficiently well over the entire domain of $\rho(\mathbf{R})$ and that can be sampled directly. Hence, as mentioned before, $S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ is chosen to be non-zero only in some domain $D(\mathbf{R}_\mathrm{i})$ of volume $\Omega(\mathbf{R}_\mathrm{i})$ around $\mathbf{R}_\mathrm{i}$, so that the proposed moves are within this domain. In that case, the condition $I(\mathbf{R}_\mathrm{f}) = I(\mathbf{R}_\mathrm{i})$ does not hold; instead $I(\mathbf{R}_\mathrm{i}) \approx S(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{i})\Omega(\mathbf{R}_\mathrm{i})$ and Eq. (14) becomes

$$\frac{A(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})}{A(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})} \approx \frac{\rho(\mathbf{R}_\mathrm{f})}{\rho(\mathbf{R}_\mathrm{i})}\frac{S(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})}{S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})}\frac{S(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{i})}{S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{f})}\frac{\Omega(\mathbf{R}_\mathrm{i})}{\Omega(\mathbf{R}_\mathrm{f})}, \tag{16}$$

from which it is apparent that the choice

$$S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = g(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) \Big/ \sqrt{\Omega(\mathbf{R}_\mathrm{f})} \tag{17}$$

yields $A(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})/A(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f}) \approx 1$, if $g(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) \overset{\propto}{\sim} \sqrt{\rho(\mathbf{R}_\mathrm{f})}$ within $D(\mathbf{R}_\mathrm{i})$ [12]. To be more precise, it can be shown[13] by Taylor expansion that if the log-derivatives of $g(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ equal those of $\sqrt{\rho(\mathbf{R}_\mathrm{f})}$ at $\mathbf{R}_\mathrm{f} = \mathbf{R}_\mathrm{i}$ then the average acceptance goes as $1 - \mathcal{O}(\Delta^m)$, where Δ is the linear dimension of $D(\mathbf{R}_\mathrm{i})$. In general, $m{=}2$, but if $D(\mathbf{R}_\mathrm{i})$ is inversion symmetric with $\mathbf{R}_\mathrm{i}$ at its center, then $m{=}3$. This is a considerable improvement compared to using a symmetric $S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ or choosing $S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) \overset{\propto}{\sim} \rho(\mathbf{R}_\mathrm{f})$ for either of which m=1.

Another possible choice, motivated by the diffusion Monte Carlo (DMC) algorithm, is

$$T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = \frac{1}{(2\pi\tau)^{3/2}} \exp\left[\frac{-(\mathbf{R}_\mathrm{f} - \mathbf{R}_\mathrm{i} - \mathbf{F}(\mathbf{R}_\mathrm{i})\tau)^2}{2\tau}\right] \tag{18}$$

where

$$F(\mathbf{R}_\mathrm{i}) = \frac{\nabla\ \Psi(\mathbf{R}_\mathrm{f})|_{\mathbf{R}_\mathrm{f}=\mathbf{R}_\mathrm{i}}}{\Psi(\mathbf{R}_\mathrm{i})}. \tag{19}$$

It has the advantage of allowing us to construct Metropolis Monte Carlo and diffusion Monte Carlo programs that share almost all the code. It can be shown that $m = 2$ for this choice of T and consequently we would expect that such an algorithm would be more efficient than one with a symmetric $S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ or one for which $S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) \overset{\propto}{\sim} \rho(\mathbf{R}_\mathrm{f})$, but less efficient than one for which $S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) \overset{\propto}{\sim} \sqrt{\rho(\mathbf{R}_\mathrm{f})}$.

The arguments of the previous two paragraphs are only rigorous in the small-step limit and so their relevance to realistic step sizes is not clear. Further, since they are based on performing Taylor expansions, they are applicable only to functions with sufficiently many derivatives – a condition that is frequently violated, at points where particles coincide, by distributions $\rho(\mathbf{R})$ of practical interest. When we consider the application of these ideas to electronic structure calculations in subsection (2.5), we will find that the ideas do in fact lead to a large reduction in the autocorrelation time provided that we employ a coordinate system such that ρ has continuous derivatives within $D(\mathbf{R}_\mathrm{i})$.

2.4. SOME OBSERVATIONS ON THE METROPOLIS METHOD

Now we make several observations as regards the Metropolis method.

1. To sample states with relative probability ρ it is not necessary to know the normalization of ρ. The Metropolis method automatically samples $\rho(\mathbf{R})/\sum\rho(\mathbf{R})$ and can only be used to calculate expectation values of the form $\sum\rho(\mathbf{R})X(\mathbf{R})/\sum\rho(\mathbf{R})$.
2. The variance of the estimate for the expectation value $\langle X\rangle$ is given by
$$\frac{T_\mathrm{corr}}{N}\left(\frac{\sum(X(\mathbf{R}_\mathrm{i}))^2}{N} - \left(\frac{\sum X(\mathbf{R}_\mathrm{i})}{N}\right)^2\right). \tag{20}$$
That is, the effective number of configurations N_eff is smaller than N by a factor of T_corr, which we define to be the autocorrelation time.
3. As mentioned before, for a given functional form of $S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$, the acceptance decreases as the volume $\Omega(\mathbf{R}_\mathrm{i})$ (and therefore the average size of the proposed moves) increases, so, there exists an optimal $\Omega(\mathbf{R}_\mathrm{i})$ for which the system evolves the fastest. There is folklore that the optimal occurs when the average acceptance ratio is close to 1/2. In fact the optimal choice may have an average acceptance that is anywhere between zero and one and in practice I have found instances where the optimum is as small as 0.2 or as large as 0.9. A much better criterion is to maximize the rate at which the system diffuses through configuration space $\langle A(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})(\mathbf{R}_\mathrm{f}-\mathbf{R}_\mathrm{i})^2\rangle$. The real measure of goodness is of course the autocorrelation time for the observable of interest.
4. If $M_1, M_2, \cdots, M_n$ are transition matrices that satisfy the stationarity condition [Eq. (5)], then we can combine these to construct a compound transition matrix M that also satisfies the stationarity condition. This is useful since one often employs elementary transition matrices M_i

that are not individually ergodic to construct a compound transition matrices M that is ergodic. Two sorts of combinations are often useful:

(a) $M = \Pi_{i=1}^{n} M_i$. Algorithmically this corresponds to performing the transitions $M_1, M_2, \cdots, M_n$ in sequence, *e.g.*, sequentially updating the Ising spins on lattice sites or the electrons in an electronic structure calculation. Note that although M satisfies the stationarity condition it need not satisfy detailed balance unless the M_i commute. (Recall that if two matrices are symmetric, their product need not be symmetric unless they commute.) This is not a cause for concern because, as noted before, detailed balance is not a necessary condition.

(b) $M = \sum_{i=1}^{n} c_i M_i$, where the c_i are non-negative constants (independent of the current state) and $\sum_{i=1}^{n} c_i = 1$. Algorithmically this corresponds to choosing the transitions M_i randomly with probabilities c_i.

5. The *heat bath* algorithm can be considered to be a special case of the (generalized) Metropolis method. There, $T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) \propto \rho(\mathbf{R}_\mathrm{f})$ for only a small set of accessible states in a domain $D(\mathbf{R}_\mathrm{i})$ in the neighborhood of $\mathbf{R}_\mathrm{i}$:

$$T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = \begin{cases} \rho(\mathbf{R}_\mathrm{f})/\sum \rho(\mathbf{R}_\mathrm{f}) & \text{if } \mathbf{R}_\mathrm{f} \in D(\mathbf{R}_\mathrm{i}) \\ 0 & \text{otherwise} \end{cases}$$

If the sum over the accessible states from $\mathbf{R}_\mathrm{i}$ and from $\mathbf{R}_\mathrm{f}$ are the same, the acceptance is unity. The heat bath algorithm is frequently used for lattice models where the normalization constant can be easily computed, but rarely for continuum systems where it is difficult to obtain the normalization constant. As an example, the heat bath algorithm can be used to update the spins in a Potts model. A single spin is updated at a time, using a proposal matrix that is proportional to the desired Boltzmann distribution conditional on all other spins being fixed. Since the distribution sampled, starting from the initial and the final states is the same, $I(\mathbf{R}_\mathrm{i}) = I(\mathbf{R}_\mathrm{f})$. So, the proposal matrix itself satisfies detailed balance, the acceptance is always unity and the system equilibrates within the subset of states accessible from $\mathbf{R}_\mathrm{i}$ in a single step. Hence, of all possible proposal matrices, T, that have the same domain $D(\mathbf{R})$, the heat bath T has the smallest autocorrelation time. This does not necessarily mean that the heat bath algorithm is the most efficient one because the computational time required to sample from the heat bath proposal matrix may be larger than for some other T. We have also ignored the possibility of devising an algorithm in which observables at neighboring Monte Carlo configurations are anticorrelated because this is rarely feasible in practice.

2.5. APPLICATION TO ELECTRONIC STRUCTURE CALCULATIONS

In order to make the ideas presented so far more concrete, we consider various choices for the proposal matrices, $T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$, used in electronic structure calculations of atoms, molecules and solids, and in particular we show how to employ the general ideas presented in subsection (2.3) to make a particularly efficient choice. This subsection and the accompanying Appendix are somewhat technical and may be skipped. For these problems, the probability distribution $\rho(\mathbf{R})$ is $|\Psi(\mathbf{R})|^2$, where $\Psi(\mathbf{R})$ is a trial wavefunction and $\mathbf{R}$ specifies the $3N$ electron coordinates. Hence Eq. (17) becomes

$$S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = |\Phi(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})| \Big/ \sqrt{\Omega(\mathbf{R}_\mathrm{f})}\,, \tag{21}$$

where $\Phi(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ is chosen such that $\Phi(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) \overset{\propto}{\sim} \Psi(\mathbf{R}_\mathrm{f})$ in $D(\mathbf{R}_\mathrm{i})$ and has log-derivatives that match those of $\Psi(\mathbf{R}_\mathrm{f})$ at $\mathbf{R}_\mathrm{f} = \mathbf{R}_\mathrm{i}$. We now discuss the three explicit forms of $S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ for which we present results here. For the first two choices $\Omega(\mathbf{R})$ is constant, independent of $\mathbf{R}$.

2.5.1. *Cartesian Coordinate Simple Metropolis*

The simplest choice is $S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ equal to a constant for $\mathbf{R}_\mathrm{f}$ within a $3N-$dimensional hypercube (box) of linear dimension 2Δ centered at $\mathbf{R}_\mathrm{i}$ and zero elsewhere. Then $T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = (2\Delta)^{-3N}$ within the box and zero elsewhere and both the S and the I factors may be omitted in Eq. (15). Aside from the question of whether one or all particles are moved in a single Monte Carlo update, this is the original Metropolis method[4] and it is the form in which it has most frequently been used despite the fact that it is less efficient than other choices.

2.5.2. *Cartesian Coordinate Directed Metropolis*

According to Eq. (21), an improved form of S is $S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = |\Phi(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})|$ within the box, if the dimension of the box is independent of $\mathbf{R}_\mathrm{i}$. The simplest choice of $\Phi(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ is a linear approximation (in each of the $3N$ cartesian coordinate directions) to $\Psi(\mathbf{R}_\mathrm{f})$ at $\mathbf{R}_\mathrm{i}$. *i.e.*

$$\begin{aligned}\Phi(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) \;=\; \prod_{k=1}^{N} & \left[1 + (x_{k,\mathrm{f}} - x_{k,\mathrm{i}}) \min\left\{\frac{1}{\Delta}, \max\left(\frac{-1}{\Delta}, \frac{\left.\frac{\partial\Psi(\mathbf{R}_\mathrm{i})}{\partial x_{k,\mathrm{f}}}\right|_{\mathbf{R}_\mathrm{f}=\mathbf{R}_\mathrm{i}}}{\Psi(\mathbf{R}_\mathrm{i})}\right)\right\}\right] \\ & \left[1 + (y_{k,\mathrm{f}} - y_{k,\mathrm{i}}) \min\left\{\frac{1}{\Delta}, \max\left(\frac{-1}{\Delta}, \frac{\left.\frac{\partial\Psi(\mathbf{R}_\mathrm{i})}{\partial y_{k,\mathrm{f}}}\right|_{\mathbf{R}_\mathrm{f}=\mathbf{R}_\mathrm{i}}}{\Psi(\mathbf{R}_\mathrm{i})}\right)\right\}\right]\end{aligned}$$

$$\left[1+(z_{k,\mathrm{f}}-z_{k,\mathrm{i}})\min\left\{\frac{1}{\Delta},\max\left(\frac{-1}{\Delta},\frac{\frac{\partial\Psi(\mathbf{R}_\mathrm{i})}{\partial z_{k,\mathrm{f}}}\Big|_{\mathbf{R}_\mathrm{f}=\mathbf{R}_\mathrm{i}}}{\Psi(\mathbf{R}_\mathrm{i})}\right)\right\}\right] \quad (22)$$

This is easier to sample than the more obvious choice,

$$\Phi(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) \;=\; 1+(\mathbf{R}_\mathrm{f}-\mathbf{R}_\mathrm{i})\cdot\frac{\nabla\Psi(\mathbf{R}_\mathrm{f})|_{\mathbf{R}_\mathrm{f}=\mathbf{R}_\mathrm{i}}}{\Psi(\mathbf{R}_\mathrm{i})} \quad (23)$$

while both forms satisfy our requirement that the log-derivatives of $\Phi(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ match those of $\Psi(\mathbf{R}_\mathrm{f})$ at $\mathbf{R}_\mathrm{f}=\mathbf{R}_\mathrm{i}$ (except at those points where the min and max functions take effect). It is possible to omit the min and max functions but in that case it is not always the case that $I(\mathbf{R}_\mathrm{i})=I(\mathbf{R}_\mathrm{f})$ and so the factors of $I(\mathbf{R}_\mathrm{i})$ and $I(\mathbf{R}_\mathrm{f})$ in Eq. (15) must be included in that case, but may be omitted if Eq. (22) is used.

Although the choice of $S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ given in Eqs. (21) and (22) is more efficient than choosing $S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ to be a constant, the improvement is not dramatic. Two features of the wavefunction restrict the size of the attempted moves. First, probable electron configurations have two core electrons at a distance of $\mathcal{O}(1/Z)$ Bohr radii from each nucleus, whereas valence electrons are typically a distance $\mathcal{O}(1)$ from the nearest nucleus[14]. Hence, if the same value of Δ is used for all the electrons, as is usually the case, then the core electrons set the length-scale for all the electrons, else a large fraction of the proposed moves are rejected. Second, $\Psi(\mathbf{R})$ has a derivative discontinuity when an electron is at a nucleus, which renders the Taylor series approximation in cartesian coordinates of $\Psi(\mathbf{R})$, used in Eq. (22), very inaccurate. (There is also a derivative discontinuity when two electrons overlap, but this is less problematic since electrons repel each other.) A natural solution to both these problems is to use spherical polar coordinates, rather than cartesian coordinates, in proposing the Monte Carlo moves, and to make the length interval, from which the radial moves of an electron are sampled, proportional to the distance of that electron from the nearest nucleus, as discussed next.

2.5.3. *Spherical Polar Coordinate Directed Metropolis*

We now describe a choice for $S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ which allows each electron to make a move appropriate to its length scale and which avoids the derivative discontinuity in $\Psi(\mathbf{R})$. We first describe a slightly simplified version of the final algorithm. Each electron moves in a volume which is the intersection of a cone which subtends an angle $2\theta_\mathrm{M}$ at the closest nucleus and a concentric spherical annulus. Let $r_{k,\mathrm{i}}$ be the initial distance of the k^{th} electron from the nucleus closest to it. The proposed move will use spherical polar coordinates centered at that nucleus. The radial moves are made in the interval $(r_{k,\mathrm{i}}/\Delta_r, r_{k,\mathrm{i}}\,\Delta_r)$. Hence the size of the radial moves is proportional to $r_{k,\mathrm{i}}$. For atoms, this ensures that if a move from the initial point to some final point can be attempted, then the reverse move can also be attempted. We will discuss the necessary minor changes for molecules and solids later. The advantages of moving in this volume are that the size of the valence electron moves is not restricted by the core electrons and that Taylor expansion of the wavefunction is valid in the region of interest since $\Psi(\mathbf{R})$ does not have derivative discontinuities in spherical polar coordinates at points where electrons and nuclei coincide.

Since $\Omega(\mathbf{R}_\mathrm{f}) = \prod_{k=1}^n (2\pi/3)(1-\cos\theta_\mathrm{M})(\Delta_r^3 - 1/\Delta_r^3) r_{k,\mathrm{f}}^3$, according to Eq. (21),

$$S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = \frac{|\Phi(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})|}{\prod_{k=1}^n r_{k,\mathrm{f}}^{3/2}}. \tag{24}$$

In order to simplify the sampling, $\Phi(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ is chosen to have the form

$$\Phi(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = \prod_{k=1}^n W(\phi_{k,\mathrm{f}}|\theta_{k,\mathrm{f}}, r_{k,\mathrm{f}}, \mathbf{R}_\mathrm{i})\ V(\theta_{k,\mathrm{f}}|r_{k,\mathrm{f}}, \mathbf{R}_\mathrm{i})\ U(r_{k,\mathrm{f}}|\mathbf{R}_\mathrm{i}). \tag{25}$$

Taking into account the $r^2 \sin\theta$ volume factor and the form of $S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ in Eq. (24), for each electron, k, the radial coordinate $r_{k,\mathrm{f}}$ is sampled from $|U(r_{k,\mathrm{f}}|\mathbf{R}_\mathrm{i})|\sqrt{r_{k,\mathrm{f}}}$, then $\theta_{k,\mathrm{f}}$ is sampled from $|V(\theta_{k,\mathrm{f}}|r_{k,\mathrm{f}}, \mathbf{R}_\mathrm{i})| \sin\theta_{k,\mathrm{f}}$ conditional upon $r_{k,\mathrm{f}}$ and finally $\phi_{k,\mathrm{f}}$ is sampled from $|W(\phi_{k,\mathrm{f}}|\theta_{k,\mathrm{f}}, r_{k,\mathrm{f}}, \mathbf{R}_\mathrm{i})|$ conditional on $r_{k,\mathrm{f}}$ and $\theta_{k,\mathrm{f}}$. The detailed form of the functions U, V, W is described in the Appendix.

It is possible to make a further small improvement by making θ_M a function $r_{k,\mathrm{f}}$ and $r_{k,\mathrm{i}}$. The reason it is particularly advantageous to do so is that most trial wavefunctions, *i.e.*, trial wavefunctions that do not have non-analytic terms in the interparticle coordinates motivated by the Fock expansion, have a finite discontinuity of magnitude Z in the local energy when two electrons approach a nucleus[17], Z being the nuclear charge. In this limit the local energy is Z Hartrees higher when the two electrons and the nucleus lie along a straight line with the electrons on opposite sides of the nucleus than with the electrons on the same side of the nucleus. It is desirable to average over this discontinuity as rapidly as possible by making large angular moves when $r_\mathrm{av} \ll 1/Z$, where $r_\mathrm{av} = (r_{k,\mathrm{i}} + r_{k,\mathrm{f}})/2$. Hence $\cos\theta_\mathrm{M}$ is chosen to be

$$\cos\theta_\mathrm{M} = \cos\theta_\mathrm{m} - \frac{1+\cos\theta_\mathrm{m}}{1+(Zr_\mathrm{av})^2}\ , \tag{26}$$

where θ_m is fixed, which has the limiting behaviors

$$\theta_\mathrm{M} \approx \begin{cases} \theta_\mathrm{m} & \text{if } r_\mathrm{av} \gg 1/Z\ , \\ \pi & \text{if } r_\mathrm{av} \ll 1/Z\ . \end{cases} \tag{27}$$

Typically, $\theta_m = \pi/2$, so that the angular moves are over a hemisphere when the electron is far from the nearest nucleus and over an entire sphere when it is close to a nucleus.

Finally, we note that although each of the above algorithms assumes that all the electrons are moved during each Monte Carlo update, it is trivial to modify the algorithms to move only one electron at each Monte Carlo step, or in fact any number in between.

2.5.4. *Generalization to Molecules and Solids*

For molecules and solids there is a minor complication which does not exist for atoms. The closest nuclei to each of the N electrons at $\mathbf{R}_i$ need not be the closest nuclei to the corresponding electrons at $\mathbf{R}_f$. For some fraction of these, the reverse move from $\mathbf{R}_f$ to $\mathbf{R}_i$ cannot be attempted, *i.e.* $S(\mathbf{R}_i|\mathbf{R}_f) = 0$, because whereas $\mathbf{R}_f$ lies in $D(\mathbf{R}_i)$, $\mathbf{R}_i$ may not lie in $D(\mathbf{R}_f)$. In that case detailed balance demands that the move from $\mathbf{R}_i$ to $\mathbf{R}_f$ be rejected *i.e.* $A(\mathbf{R}_f|\mathbf{R}_i) = 0$. Since the number of such moves is small and since these rejections may be performed on purely geometrical grounds, (they do not require calculating the wavefunction or its derivatives at $\mathbf{R}_f$), this does not result in an appreciable loss of efficiency.

2.5.5. *Results for Atoms and Molecules*

In Fig. 1 we show the acceptance of the one-electron-move versions of the algorithms presented in subsections 2.5.1, 2.5.2 and 2.5.3 as a function of the initial distance of the electron from the nucleus using a simple Neon atom wavefunction. As expected, the acceptance of the cartesian coordinate directed Metropolis method is always greater than that of the cartesian coordinate simple Metropolis method. However, the acceptance of both these algorithms becomes very small near the nucleus for the reasons described earlier. In contrast the spherical-polar coordinate directed Metropolis method has unit acceptance at the nucleus and a large acceptance in all regions of space. The acceptance is smaller than that of the cartesian coordinate directed Metropolis method in the region $0.4 < r < 4$ but it is in fact much more efficient even in this region of space because the proposed moves are very much larger. Note $\Delta_r = 5$ and $\theta_m = \pi/2$ were used.

The efficiency of the algorithms is inversely proportional to the autocorrelation time of observables of interest. Table 1 shows the autocorrelation time of the energy for four algorithms and four wavefunctions. The four algorithms are: 1) the cartesian coordinate simple Metropolis algorithm moving all electrons at each Monte Carlo step, 2) the cartesian coordinate directed Metropolis algorithm moving all electrons at each Monte Carlo step, 3) the cartesian coordinate directed Metropolis algorithm moving only one electron at each Monte Carlo step and 4) the spherical-polar coordi-

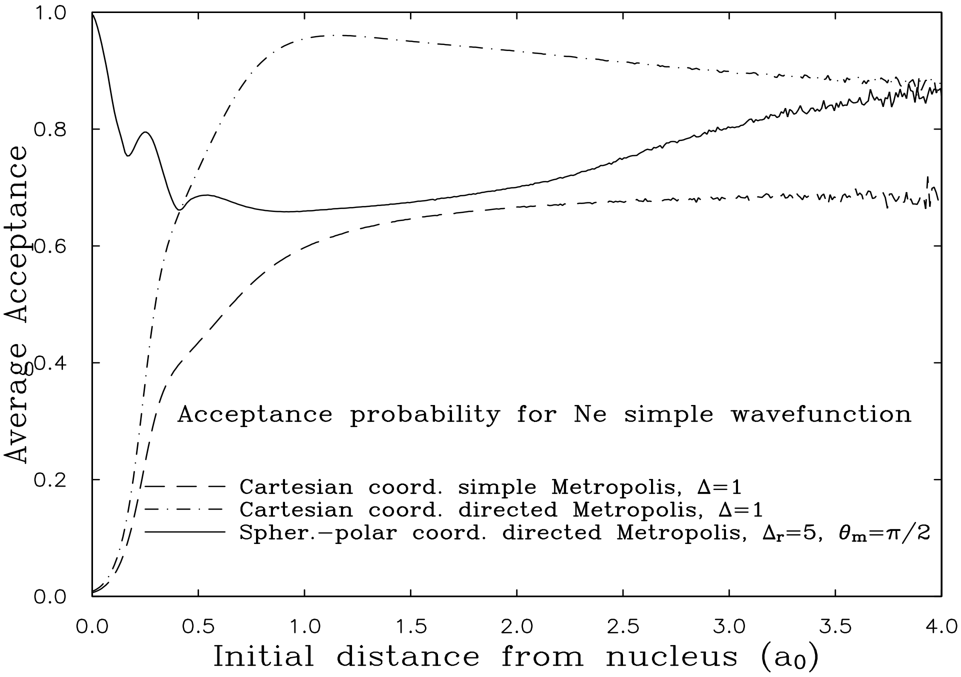

Figure 1. Acceptance as a function of the initial distance of the electron from the nucleus using a simple Neon atom wavefunction for the three single-electron-move algorithms described in the text.

nate directed Metropolis algorithm moving only one electron at each Monte Carlo step. The four wavefunctions used are: 1) a simple Ne wavefunction, 2) a good Ne wavefunction, 3) a simple Ar wavefunction and 4) a simple Li_2 wavefunction. The simple wavefunctions consist of a determinant multiplied by a simple two-body (electron-electron) Jastrow function, whereas the good wavefunction consists of a determinant multiplied by a three-body (electron-electron-nucleus) Jastrow function.

The measure of efficiency of the algorithm is the autocorrelation time, which is determined as follows. The entire Monte Carlo run, (after discarding the equilibration updates), consists of N Monte Carlo updates that are divided into $N_{\rm b}$ blocks each consisting of $N_{\rm s}$ Monte Carlo steps for each of the N electrons. The local energy is measured after each Monte Carlo update. The autocorrelation time is given by $T_{\rm corr} = N_{\rm s}(\sigma_{\rm b}/\sigma)^2$, where σ and $\sigma_{\rm b}$ are the root-mean-square fluctuations of the individual energies and of the block average energies respectively. $N_{\rm s}$ must be chosen such that $N_{\rm s} \gg T_{\rm corr}$. The autocorrelation times presented in Table 1 were obtained using values of $N_{\rm s}$ that were at least 100 times greater than $T_{\rm corr}$. Using $N_{\rm s} \approx 10\, T_{\rm corr}$ results in estimates of $T_{\rm corr}$ that are too low by as much as

TABLE 1. Autocorrelation times for the four wavefunctions and the four algorithms discussed in the text. The four algorithms are: 1) cartesian coordinate simple Metopolis with all-electron moves, 2) cartesian coordinate directed Metopolis with all-electron moves, 3) cartesian coordinate directed Metopolis with one-electron moves and 4) polar coordinate directed Metopolis with one-electron moves, In order to have a fair comparison we give in the last column $T^*_{\mathrm{corr}} = T_{\mathrm{corr}}$ for algorithms 1 and 2, and $T^*_{\mathrm{corr}} = 2T_{\mathrm{corr}}$ for algorithms 3 and 4. In algorithm 4, $\theta_{\mathrm{m}} = \pi/2$ for Ne and Ar and $\theta_{\mathrm{m}} = \pi$ for Li_2. $\bar{A}$ is the average acceptance. The uncertainty in T^*_{corr} is typically 10% of its value. (Taken from Ref. [10].)

Wavefunction	Algorithm	Δ	Δ_r	$\bar{A}$	T^*_{corr}
Ne simple	1	0.25	–	0.288	84.
$\bar{E} = -128.716$ H	2	0.3	–	0.661	28.
$E_{\mathrm{corr}} = 43\%$	3	0.8	–	0.769	13.
$\sigma = 1.8$ H	4	–	5.	0.708	2.0
Ne good	1	0.25	–	0.290	37.
$\bar{E} = -128.901$ H	2	0.3	–	0.663	11.
$E_{\mathrm{corr}} = 91\%$	3	0.8	–	0.771	7.2
$\sigma = 0.91$ H	4	–	5.	0.708	1.7
Ar simple	1	0.12	–	0.307	190.
$\bar{E} = -527.1$ H	2	0.2	–	0.474	44.
$E_{\mathrm{corr}} = 37\%$	3	0.5	–	0.813	13.
$\sigma = 4.0$ H	4	–	5.	0.620	2.2
Li_2 simple	1	0.75	–	0.268	210.
$\bar{E} = -14.9476$ H	2	1.	–	0.614	56.
$E_{\mathrm{corr}} = 61\%$	3	2.	–	0.616	32.
$\sigma = 0.41$ H	4	–	8.	0.775	5.8

20%. When Monte Carlo moves consist of moving one electron at a time, it takes twice the computer time to move all the electrons as compared to an algorithm where they are all moved at once [2]. Hence, in order to have a fair comparison, Table 1 has values of $T^*_{\mathrm{corr}} = 2T_{\mathrm{corr}}$ for algorithms 3 and 4. We note that σ and the variational energy $\bar{E}$ depend on the trial wavefunction but not on the algorithm used. T_{corr} and the acceptance $\bar{A}$ depend on both the algorithm and the trial wavefunction.

For each of the algorithms the values of Δ, or, Δ_r and θ_{m} were crudely optimized to yield the smallest possible T_{corr}. Table 1 shows the values of T^*_{corr} for approximately optimal values of the parameters. Moderate variations of the parameters about their optimal values affect the efficiency of the algorithm only slightly. For example, changing the value of Δ_r from 5 to 4 alters the value of T_{corr} by less than 10% for each of the wavefunctions. We observe that large improvements in T_{corr} are achieved. For the four wavefunctions, the autocorrelation times get smaller by factors of 42, 22, 86 and 36 respectively as we progress from algorithm 1 to 4 and by factors of 6.5, 4.2, 5.9 and 5.5 in going from algorithm 3 to 4.

Note that for each of the four algorithms T_{corr} is smaller for the good Ne wavefunction than for the simple Ne wavefunction. The likely reason for this is that the good Ne wavefunction has a more rapidly varying local energy (but with a smaller amplitude of course) and consequently it takes fewer Monte Carlo steps to wander from a region where the local energy is too high to one where it is too low and vice versa. Hence the gain in efficiency from improving the wavefunction is greater than would be supposed by merely comparing their respective values of σ.

The only difference between algorithms 2 and 3 is that in the former Monte Carlo moves consist of moving all electrons whereas in the latter only one electron is moved per Monte Carlo move. Comparison of the results for these algorithms shows, as has been noted before [2], that for systems with many electrons, and for this class of algorithms, it is more efficient to move one electron at a time rather than all at once.

For algorithm 1 there is a considerable increase in T_{corr} in going from Ne to Ar. On the other hand for algorithm 4, T_{corr} increases very little. This does not mean that calculations of heavy atoms can be performed as rapidly as those of light atoms, unless, of course, pseudopotentials are used to eliminate the core electrons. First, the time for evaluating determinants in the wavefunctions scales as N^3. Second, the fluctuation in the local energy σ increases with the number of electrons N. The dependence of σ on N is different depending on whether one is interested in progressively heavier neutral atoms, or in increasing the number of atoms of the same species. It is not known how σ scales with the nuclear charge Z. In subsection (4.3) we show empirically that going from the lightest to the heaviest first-row homonuclear molecules the increase is somewhat larger than Z^2, but that for the heavier molecules the increase is nearly linear in Z. The statistical error goes down as the square root of the number of Monte Carlo updates or equivalently the computer time. Hence, if we assume $\sigma \sim Z^2$, then the computer time needed to obtain a given statistical error scales as $T \sim Z^7$, while if $\sigma \sim Z$, then $T \sim Z^5$. On the other hand, the scaling is not as bad if one is interested in increasingly larger clusters of atoms of fixed Z. In that case $\sigma \sim \sqrt{N}$ and so $T \sim N^4$. Further, if one is interested in achieving a constant statistical error *per atom*, then $T \sim N^2$. The scaling for each of these situations is worse of course if one uses a Metropolis method for which $T_{\rm corr}$ increases with system size. The poor scaling with respect to Z is the reason why it is necessary to employ pseudopotentials in calculations involving second row and heavier atoms (see chapters by Fahy and Mitas in this volume).

From Table 1 we observe that $T_{\rm corr} = T^*_{\rm corr}/2 = 0.85$ for the good Ne wavefunction using the spherical-polar coordinate directed Metropolis algorithm. This may at first sight seem impossible since one may suppose that the best possible value for $T_{\rm corr}$ is 1. The reason it is in fact possible is that we are using a single-electron move algorithm and updating the local energy at each Monte Carlo step but $T_{\rm corr}$ is in units of all-electron moves.

3. Variance Minimization Method for Optimizing Wavefunctions

Optimized trial wavefunctions that closely approximate eigenstates of Hamiltonians are essential ingredients of accurate electronic structure calculations employing quantum Monte Carlo methods. The quality of these trial wavefunctions is relevant both for expectation values of physical interest and for the variance of the Monte Carlo estimators. As the trial wave function approaches an exact eigenstate of the Hamiltonian, the energy and expectation values of operators commuting with the Hamiltonian satisfy a zero-variance principle, *i.e.*, the expectation values approach the exact eigenstate values, while the Monte Carlo variance goes to zero.

A major advantage of quantum Monte Carlo methods is that there is no restriction on the form of trial wavefunction $\Psi(\mathbf{R})$. Hence any insight one may have, as regards the nature of the many-body correlations, can be built into $\Psi(\mathbf{R})$ and tested. To exploit this freedom it is necessary to have a method for optimizing arbitrary wavefunctions.

The method consists of the minimization of the variance of the local energy [15, 16] over a set of N_c configurations $\{\mathbf{R}_i\}$ sampled from the square of the best wavefunction, Ψ_0, available before we start the optimization, *i.e.*,

$$\sigma^2_{\rm opt}[\Psi] = \sum_i^{N_c} \left[\frac{\mathcal{H}\Psi(\mathbf{R}_i)}{\Psi(\mathbf{R}_i)} - \bar{E}\right]^2 w(\mathbf{R}_i) \Big/ \sum_i^{N_c} w(\mathbf{R}_i), \tag{28}$$

where

$$\bar{E} = \sum_i^{N_c} \frac{\mathcal{H}\Psi(\mathbf{R}_i)}{\Psi(\mathbf{R}_i)} w(\mathbf{R}_i) \Big/ \sum_i^{N_c} w(\mathbf{R}_i), \tag{29}$$

is the average energy over the sample of configurations. The weights $w(\mathbf{R}_i) = |\Psi(\mathbf{R}_i)/\Psi_0(\mathbf{R}_i)|^2$ are introduced to allow the nodes of the wavefunction to move freely during the optimization and to provide the correct weighting.

The ratio of each of the weights to the average weight is not allowed to exceed a maximum value; otherwise, it is possible for the optimizer to achieve a small value of $\sigma_{\rm opt}$ by having a few configurations gain a very large weight and a local energy that is very close to $\bar{E}$. Typically, 3000 configurations are sufficient to optimize a wavefunction with 50-100 parameters.

Variance minimization is more effective than energy minimization because the smallest possible variance is zero whereas the energy is unbounded from below on the necessarily sparse set of N_c configurations on which the optimization is performed. (In a many-dimensional space any number of configurations, that can fit in computer memory, is a sparse sample.) A direct attempt to minimize the energy of a many-dimensional, multi-parameter wavefunction frequently leads to a wavefunction that has a low estimated energy on the N_c configurations used to perform the optimization but a higher energy than that of the initial wavefunction on a new set of Monte Carlo configurations. The "optimization" results in a worse wavefunction!

There are other advantages to variance minimization. Since the function to be minimized is a sum of squares, effective optimization techniques such as the Levenberg-Marquardt method can be employed. In practice we employ a modified [18] Levenberg-Marquardt method wherein up-hill moves are permitted under certain conditions. Another advantage is that, since any eigenstate has zero variance, the method can be used to optimize trial wavefunctions for excited states as well, including those that have the same symmetry as the ground state. Finally, the method can be used even for Hamiltonians that have a spectrum that is unbounded from below, *e.g.*, a relativistic Hamiltonian.

In practice, one optimizes a combination of the variance and the energy by a trivial modification of the procedure described above. One possibility [16] is to replace $\bar{E}$ by $E_{\rm guess}$, where $E_{\rm guess}$ is chosen to be slightly lower than the best current estimate for the energy. A more elegant solution [19] is to optimize $\sigma^2/\bar{E}^2$.

As a final remark we note that the variance minimization method has been extended to simultaneously optimize several trial states [20], to find optimized trial states of transfer matrices [21] and to find the subdominant trial state of a Markov matrix [22]. The latter two are of relevance to static and dynamic equilibrium problems in statistical mechanics.

4. Forms of Many-body Wavefunctions

We next discuss the functional form of the wavefunctions used in electronic structure calculations of atoms and molecules. Quantum chemistry methods such as Configuration Interaction (CI) expand the wavefunction as a linear combination of determinants of single-particle orbitals. In the limit of an infinite sum this is exact but the rate of convergence is very slow, in part because the true wavefunction has a cusp at electron coincidences but the determinants, used to expand the wavefunction, do not. In the quantum Monte Carlo methods the above form of the wavefunction is multiplied by a generalized Jastrow factor (also known as a Feenberg factor), expressed in terms of electron-electron and electron-nucleus coordinates, which allows the trial wavefunction to have the correct cusp at electron coincidences. Since the Jastrow factor is everywhere positive, it does not alter the nodes of the wavefunction and so it is necessary to have multiple determinants to accurately describe the nodes of the wavefunction. The intuition one has is that the near-degeneracy correlation is most effectively described by the linear combination of determinants, whereas the Jastrow factor describes well the remaining correlation (which requires a very large number of determinants in a purely determinantal wavefunction) and allows the imposition of the electron-electron cusps. In practice it is possible to construct rather compact quantum Monte Carlo wavefunctions with typically 50 to 100 parameters that have an energy that is comparable to a CI wavefunction with orders of magnitude more parameters.

The trial wavefunctions, commonly used in quantum Monte Carlo applications to problems in electronic structure, are in fact sums of products of up- and down-spin determinants of single-particle orbitals multiplied by a Jastrow factor

$$\Psi = J \sum_i d_i D_i^{\uparrow} D_i^{\downarrow}. \tag{30}$$

$D_i^{\uparrow}$ and $D_i^{\downarrow}$ are the Slater determinants of single particle orbitals for the up- and down-spin electrons respectively. For atoms and molecules, the orbitals are usually expanded in linear combinations of products of Slater basis functions and spherical harmonics centered at the nuclei, whereas for pseudopotential calculations of solids a plane-wave expansion is usually used.

A natural question to ask is: "why is it acceptable to use a wavefunction that is not totally antisymmetric under the exchange of electronic coordinates, such as that in Eq. (30)?" A wavefunction Ψ for an N-electron system ($N = N_\uparrow + N_\downarrow$) in an $S_z = (N_\uparrow - N_\downarrow)/2$ state can be decomposed in terms of its spin components as

$$\begin{aligned}\Psi(\mathbf{r_1} s_1, \cdots, \mathbf{r_N} s_N) &= \sum_{i=1}^{K} F_i(\mathbf{r_1}, \cdots, \mathbf{r_N})\, \zeta_i(s_1, \cdots, s_N) \\ &= \mathcal{A}\left\{F_1(\mathbf{r_1}, \cdots, \mathbf{r_N})\, \zeta_1(s_1, \cdots, s_N)\right\}, \end{aligned} \tag{31}$$

where

$$\zeta_1(s_1, ...s_N) \;=\; \chi_\uparrow(s_1)...\chi_\uparrow(s_{N_\uparrow})\chi_\downarrow(s_{N_\uparrow+1})...\chi_\downarrow(s_N). \tag{32}$$

The sum in Eq. (31) extends over the $K = \binom{N}{N_\uparrow} = N!/(N_\uparrow! N_\downarrow!)$ spin functions (or spin assignments) ζ_i generated by permuting particle indices in ζ_1, and $\mathcal{A}$ is the total antisymmetrizer. It follows from the antisymmetry of Ψ under the interchange of particle indices that each function F_i is antisymmetric under the interchange of like-spin electrons and that the F_i are all the same except for a relabeling of the particle indices and a change in sign for odd permutations. Since the spin functions $\zeta_i(s_1, ..., s_N)$ form an orthonormal set, *i.e.*,

$$\sum_{s_1,...,s_N} \zeta_i(s_1, ..., s_N)\zeta_j(s_1, ..., s_N) = \delta_{i,j}, \tag{33}$$

the expectation value of an operator that is independent of spin is the same if we use the fully antisymmetric wavefunction Ψ or just one spatial function, say

$$F_1(\mathbf{r_1}, \cdots, \mathbf{r_N}) \;=\; \sum_i d_i \mathbf{D}_i^\uparrow(\mathbf{r_1}, \cdots, \mathbf{r_{N_\uparrow}})\mathbf{D}_i^\downarrow(\mathbf{r}_{N_\uparrow+1}, \cdots, \mathbf{r}_N), \tag{34}$$

which justifies the functional form given in Eq. (30).

4.1. SPIN-CONTAMINATION

The wavefunction of Eq. (30) is one term in the complete wavefunction, which can be written as a product of functions Φ and J that are, respectively, antisymmetric and symmetric under particle exchange. Expanding each of Φ and J in terms of its spin components, we obtain

$$\begin{aligned}
\Psi(\mathbf{r_1}s_1, \cdots, \mathbf{r_N}s_N) &= \Phi(\mathbf{r_1}s_1, \cdots, \mathbf{r_N}s_N)\; J(\mathbf{r_1}s_1, \cdots, \mathbf{r_N}s_N) \\
&= \left(\sum_{i=1}^{K} f_i(\mathbf{r_1}, \cdots, \mathbf{r_N})\; \zeta_i(s_1, \cdots, s_N)\right)\left(\sum_{i=1}^{K} G_i(\mathbf{r_1}, \cdots, \mathbf{r_N})\; \zeta_i(s_1, \cdots, s_N)\right) \\
&= \sum_{i=1}^{K} F_i(\mathbf{r_1}, \cdots, \mathbf{r_N})\; \zeta_i(s_1, \cdots, s_N),
\end{aligned} \tag{35}$$

where $F_i = f_i G_i$ and we have used the relation $\zeta_i\zeta_j = \delta_{i,j}\zeta_i$.

The determinantal part of the trial wavefunction is usually chosen to have the same symmetry as the eigenstate it approximates, in particular it is an eigenstate of $\hat{S}^2$ and $\hat{S}_z$. It is shown in Ref. [23] that it is an eigenstate of S^2 if the functions f_i in Eq. (35) satisfy certain linear relations.

If the Jastrow factor is symmetric under the exchange of just the spatial coordinates, the F_i satisfy the same linear relations as the f_i and the above wavefunction is also an eigenstate of $\hat{S}^2$. However, for reasons explained in the next subsection the Jastrow factor usually does not usually not have this symmetry and consequently the wavefunction may not be an eigenstate of S^2, *i.e.*, it is *spin contaminated.* However, it is found that for well optimized wavefunctions the spin contamination is very small, typically less than 10^{-3} for light atoms [23].

4.2. CUSP CONDITIONS

In order to obtain a wavefunction with a small variance of the local energy, it is important that the trial wavefunctions satisfy a set of *cusp conditions* which prescribe the proper derivative discontinuity at the collision points to ensure that the divergence in the local potential is canceled by an opposite divergence in the local kinetic energy. Kato[24] first rigorously derived these conditions as a general property of Coulombic systems. He showed that in the limit that two particles of masses m_i and m_j and charges q_i and q_j approach each other and all other interparticle distances remain larger than zero,

$$\left.\frac{\partial \hat{\Psi}}{\partial r_{ij}}\right|_{r_{ij}=0} = \mu_{ij} q_i q_j \Psi(r_{ij}=0), \tag{36}$$

where $\mu_{ij} = m_i m_j/(m_i+m_j)$ is the reduced mass of the two-particle subsystem and $\hat{\Psi}$ is the average of Ψ over an infinitesimally small sphere centered at $r_{ij} = 0$. Pack and Byers-Brown[25] generalized Kato's result by solving the multiparticle Schrödinger equation in the neighborhood of a two-particle coincidence. Writing the wavefunction near such a collision point, $\mathbf{r}_{ij} = 0$, as

$$\Psi = \sum_{l=l_0}^{\infty} \sum_{m=-l}^{l} f_{lm}(r)\, r^l\, Y_{lm}(\theta, \phi), \tag{37}$$

where $r = r_{ij},\ \theta = \theta_{ij},\ \phi = \phi_{ij}$, and the $Y_{lm}(\theta, \phi)$ are the usual spherical harmonics and, expanding $f_{lm}(r) = \sum_{k=0}^{\infty} f_{lm}^{(k)} r^k$, they found for electron-electron collisions,

$$f_{lm}(r) = f_{lm}^{(0)} \left[1 + \frac{1}{2(l+1)} r + O(r^2)\right]. \tag{38}$$

The terms $r^l\, Y_{lm}(\theta, \phi)$ in Eq. (37) are analytic functions that can be well described by the determinants appearing in Eq. (30) while the terms

$[1 + \frac{1}{2(l+1)}r]$ are nonanalytic functions of the individual electron coordinates, having cusps of magnitude $\frac{1}{2(l+1)}$ (though they are of course analytic functions of the interparticle distance r). The local energy at the collision point is finite provided that the cusp condition [Eq. (38)] is obeyed for the lowest value l_0 appearing in the sum in Eq. (37). In the case of parallel-spin electrons $l_0 = 1$ and the wavefunction has a cusp of 1/4 at the coincidence point. In the case of antiparallel-spin electrons there are three possibilities, $l_0 = 0, 1, 2$ corresponding to cusps of 1/2, 1/4 and 1/6. States with $l_0 = 1$ may be avoided by always choosing, for a given value of S, a state with extremal S_z, *i.e.*, $S_z = \pm S$. (The S_z value of a state depends on the arbitrary choice of a coordinate system and therefore one is free to choose a coordinate system in which $S_z = \pm S$.) Unlike the $l_0 = 1$ case, the $l_0 = 2$ case cannot be circumvented by considering a different state with the same expectation values. However, it occurs only rarely for ground states [26] and consequently the usual prescription of using a Jastrow factor with a cusp of 1/4 (1/2) for parallel- (antiparallel)-spin electrons is correct for most ground state wavefunctions, for $S_z = \pm S$ states. Since the Jastrow factors for parallel- and antiparallel-spin electrons are different, quantum Monte Carlo wavefunctions typically have a small spin-contamination, as explained in the previous subsection.

4.3. RESULTS FOR MULTIDETERMINANTAL WAVEFUNCTIONS

Improvements to quantum Monte Carlo wavefunctions, while staying within the form given in Eq. (30), can be made by either adding more determinants or by including higher-order correlations in the Jastrow factor. Unfortunately there has not been sufficient systematic study of either, in part because, in implementing either of these improvements, multiple choices have to be made in choosing the precise functional form and also because one is never completely sure that the best parameters, for a given functional form, have been obtained. Nevertheless, some conclusions can be drawn from existing studies and in this subsection we review a study of the improvements obtained by adding low-lying determinants [27] to wavefunctions of first-row homonuclear diatomic molecules and in the next subsection we review the improvements obtained from using higher-order correlations [29] in the Jastrow factor.

Figure **??** shows the total energies obtained in variational Monte Carlo and diffusion Monte Carlo and the root-mean-square fluctuations of the local energy, σ_{VMC} and its autocorrelation time T_{corr} in variational Monte Carlo. Results are presented for both single-determinant and multi-determinant wavefunctions [27]. The multiconfiguration wavefunctions have 2 to 6 configuration state functions, corresponding to 2 to 16 determi-

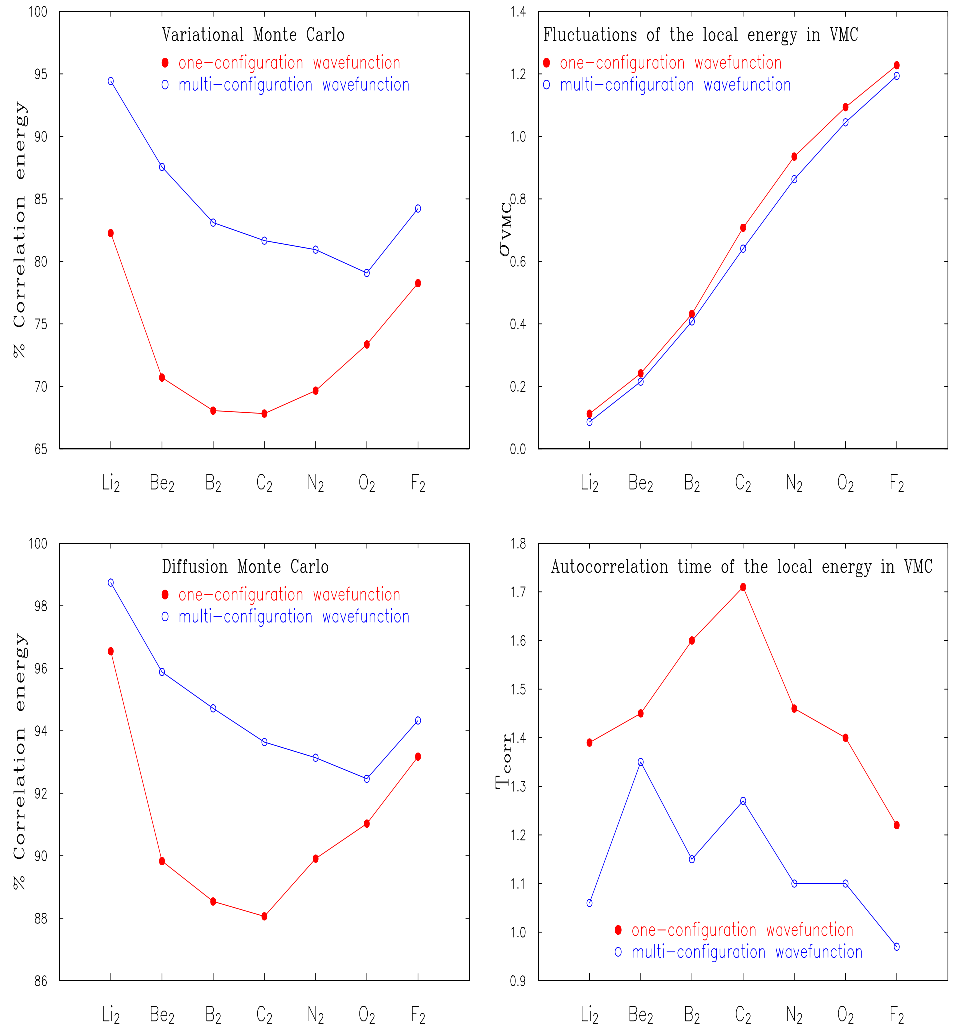

Figure 2. Percentage of correlation energy in VMC (upper left) and DMC (lower left), root mean square fluctuation σ_{VMC} (upper right) and autocorrelation time of the local energy in VMC (lower right) with one-configuration (red, full circle) and multi-configuration (blue, empty circle) wavefunctions. (Taken from Ref. [27].)

nants. The number of configurations/determinants is very small compared to what is normally used in Configuration Interaction (CI) calculations. The percentage of correlation energy recovered ranges from 68-82% for single-configuration and 79-94% for multi-configuration wavefunctions in variational Monte Carlo. The corresponding numbers are 88.1-96.5% and 92.5-98.7% in diffusion Monte Carlo. For the single-configuration wavefunctions, the smallest percentage of correlation energy recovered is not for the heaviest molecules but rather for the molecules in the middle of the row owing to the strong multi-configurational nature of their true ground state. For the multi-configuration wavefunctions, the smallest percentage moves more to the right of the periodic table where, possibly, excitations to the next shell become important.

There is considerable resemblance in the shapes of the variational Monte Carlo and diffusion Monte Carlo curves and the multi-configuration energies are consistently better than their single-configuration counterparts. Therefore, it is clear that when more configurations are added, not only do $\sigma_{\rm VMC}$ and $\rm E_{VMC}$ improve but there is also an improvement in the shape of the nodal surface, flaws in which limit the accuracy of the fixed-node diffusion Monte Carlo energies. The autocorrelation time $T_{\rm corr}$ is also considerably smaller for the multiconfiguration wavefunctions. The likely reason for this is that when the wavefunction is improved, not only is the magnitude of the fluctuations of the local energy reduced (this would leave $T_{\rm corr}$ unchanged) but the local energy changes more rapidly in configuration space and consequently electrons need travel a shorter distance in order for the local energy to decorrelate.

These results shed some light on the question of how the computational complexity of quantum Monte Carlo scales with atomic number Z. In going from $\rm Li_2$ to $\rm F_2$, the root-mean-square fluctuation of the local energy, $\sigma_{\rm VMC}$, increases by factors of 11 and 13 for the single and the multi-configuration wavefunctions respectively, indicating a somewhat faster than Z^2 scaling of $\sigma_{\rm VMC}$ with Z. However, for the heaviest molecules the scaling is very nearly linear in Z. The scaling of $\sigma_{\rm VMC}$ with Z should be taken into account in figuring the scaling of the computational cost of QMC methods with atomic number. In order to better estimate this scaling, it would be useful to have results on some second-row homonuclear diatomic molecules as well.

4.4. EFFECT OF FOUR-BODY CORRELATION IN JASTROW FACTORS

Early quantum Monte Carlo work was done using a a two-body (electron-electron) Jastrow factor. Since then it has become apparent that inclusion of three-body (electron-electron-nucleus) terms results in a large reduction of the fluctuations of the local energy and a large lowering of the energy [16]. The results for molecules presented in the previous subsection were all obtained with a three-body Jastrow factor. It is natural to ask whether it is worthwhile to include four-body ($\rm e^3$-n) correlations in the Jastrow factors. A systematic expansion of the Jastrow factor consists of a polynomial in scaled inter-particle coordinates. By “scaled” we mean that a monotonic function is used to map the infinite range of the inter-particle coordinates onto a finite interval. As shown in Table 2 the number of terms in such an expansion grows rapidly, both as a function of the body order and as a function of the polynomial order of the expansion. If some or all of the particles are identical then the theory of invariants can be used to reduce the computational effort required to evaluate the function. This was first done for identical bosons, for which it was possible to include upto 5-body correlations [28] and later for electronic structure calculations where four-body (e^3-n) correlations were included [29].

TABLE 2. Number of monomials of degree p in d inter-particle distances correlating n particles. (Taken from Ref. [29].)

		Number of terms of polynomial order					
n	d	$p=1$	$p=2$	$p=3$	$p=4$	$p=5$	Total
3	3	3	6	10	15	21	55
4	6	6	21	56	126	252	461
5	10	10	55	220	715	2002	3002
6	15	15	120	680	3060	11628	15503

In Table 3, we show the variational Monte Carlo energies E^{VMC} and the standard deviation of the local energies σ_{VMC} of the Li, Be, and Ne atoms for wavefunctions containing two-body, three-body and four-body correlations. Inclusion of the three-body correlations results in a large improvement in the energy and a large reduction in the standard deviation of the local energy. For Li, Be and Ne, 95%, 89% and 84% of the part of the correlation energy missing in the two-body wavefunctions is recovered while the standard deviation of the local energy is reduced by factors of 6, 4 and 2 respectively. Although not shown in the table, a further advantage accrues from the fact that the auto-correlation time of the local energies sampled in variational Monte Carlo and in diffusion Monte Carlo are somewhat reduced. We observe from Table 3 that inclusion of the four-body correlations also results in a significant but smaller improvement in both the variational energies and the fluctuations of the local energy. The improvement gets smaller with increasing atomic number and is disappointingly small for Ne.

Also shown in Table 3, are the diffusion Monte Carlo energies. Since the nodes of the wavefunctions are determined by the determinantal part of the wavefunctions only, it is clear that if the same determinants are used in two wave functions, the fixed-node diffusion Monte Carlo energies must be the same to within statistical error. However, the determinantal part of the

TABLE 3. Total energies of Li, Be, and Ne atoms obtained in variational and diffusion Monte Carlo. For each atom, the first, second, and third rows respectively contain results using wavefunctions with n-body correlations respectively with $n = 2$, $n = 3$, and $n = 4$. The number of configuration state functions in the determinantal part is one for Li and Ne and two for Be. The "exact" total energy from Ref.[30] is -7.47806, -14.66736 and -128.9376 H, for Li, Be and Ne respectively, while the corresponding correlation energies are -0.04533, -0.09434 and -0.3905 H. $E_c^{\rm VMC}$ and $E_c^{\rm DMC}$ are the percentages of correlation energy recovered in variational and diffusion Monte Carlo. $\sigma_{\rm VMC}$ is the root mean square fluctuation of the local energy in variational Monte Carlo. The numbers in parentheses are the statistical errors in the last digit. The last column is the efficiency of the four-body correlations as defined in Eq. (39). Energies are in Hartree atomic units. (Taken from Ref. [29] and modified.)

Atom	n	$E^{\rm VMC}$	$E_c^{\rm VMC}$ (%)	$E^{\rm DMC}$	$E_c^{\rm DMC}$ (%)	$\sigma_{\rm VMC}$	η
Li	2	-7.47427(4)	91.6	-7.47801(3)	99.9	0.24	
	3	-7.47788(1)	99.6	-7.47803(1)	99.9	0.037	95%
	4	-7.47797(1)	99.8	-7.47803(1)	99.9	0.028	60%
Be	2	-14.66088(5)	93.1	-14.66689(4)	99.5	0.35	
	3	-14.66662(1)	99.2	-14.66723(1)	99.9	0.089	89%
	4	-14.66681(1)	99.4	-14.66726(1)	99.9	0.078	31%
Ne	2	-128.713(2)	42.5	-128.919(2)	95.2	1.9	
	3	-128.9008(1)	90.6	-128.9242(1)	96.6	0.90	84%
	4	-128.9029(3)	91.1	-128.9243(8)	96.6	0.88	9%

wavefunctions were reoptimized when the higher-body terms in the Jastrow exponents were added, since one hopes that the additional freedom would allow the determinantal part of the wavefunction to attain a more optimal nodal structure. However, only a small improvement in the diffusion Monte Carlo energy is found upon going from the two-body to the three-body Jastrow, while the difference of the three-body and four-body cases is within the statistical error. The reader should not conclude from these results that, in diffusion Monte Carlo, improvements in the wavefunction resulting from improved Jastrow factors are worthless. First of all, improved Jastrow factors are likely to yield more reliable expectation values of operators that do not commute with the Hamiltonian. Secondly, even if one is interested only in the diffusion Monte Carlo energy, the improved Jastrow factor results in smaller statistical errors for a given number of Monte Carlo steps and usually also in smaller time-step errors, leading to a more reliable extrapolation to the zero time-step limit. The time-step error is much smaller for the three-body Jastrow wavefunction than for the corresponding two-body

Jastrow wavefunction. For example, the three-body Jastrow wavefunction for Be has a time-step error that is approximately 30 times smaller than that of the two-body Jastrow wavefunction. However, the time-step error is of comparable magnitude for the three-body and four-body Jastrow wavefunctions.

The best possible Jastrow factor (one that includes all-body correlations and has an infinite-order polynomial) would have a variational Monte Carlo energy that equals the fixed-node diffusion Monte Carlo energy. Hence, a measure of the efficiency of the four-body contributions to the Jastrow factor is

$$\eta = \frac{E_{\mathrm{VMC}}^{(n-1)-\mathrm{body}} - E_{\mathrm{VMC}}^{n-\mathrm{body}}}{E_{\mathrm{VMC}}^{(n-1)-\mathrm{body}} - E_{\mathrm{DMC}}}. \tag{39}$$

The percent efficiency is much larger going from two-body correlations to three-body correlations than going from three-body correlations to four-body correlations. In the latter case, it rapidly decreases from 60% for Li to 31% for Be to 9% for Ne. The reduction in the root mean square fluctuations of the local energy is 24% for Li, 12% for Be and 2% for Ne. These somewhat disappointing results are probably due to flaws of the nodal surface of the trial wavefunction, which the four-body interaction Jastrow factor is not designed to correct. That is, the approximate, fixed-node, diffusion Monte Carlo wavefunction has discontinuous derivatives almost everywhere across the nodal surface. On the other hand, a Jastrow factor expressed in terms of inter-particle coordinates, will have non-analyticities only at the $(3N-3)$-dimensional surface where particles coincide, which constitutes only a vanishingly small fraction of the entire $(3N-1)$-dimensional nodal surface. Hence, over most of the nodal surface, we are attempting to describe a non-analytic function as a finite sum of analytic functions and we expect the convergence to be slow. The validity of the above argument has been tested by performing calculations for the nodeless, bosonic ground states of the Hamiltonians of the same atoms. As expected, the efficiency for the bosonic states is very much greater than for the fermionic states and the reader is referred to Ref. ([29]) for details.

Another interesting observation that emerges from the calculations on the bosonic states is that the fluctuations of the local energy σ are much smaller than for the corresponding fermionic states, despite the fact that the absolute energies are larger for the bosonic states. This indicates that the difficulty of obtaining a good parametrization of the nodal surface of a fermionic state, or more generally of any excited state, is a significant bottleneck in the search for optimized trial wavefunctions.

For any given wavefunction it is relevant to ask whether further improvements can be made most economically by increasing the variational freedom of the determinantal part of the wavefunction (by increasing the number of single-particle basis functions and/or the number of determinants) or by improving the Jastrow part of the wavefunction (by increasing the polynomial-order and/or the body-order). It is clear that if we follow the former route to the limit of a complete basis of Slater determinants then the exact result can be obtained, even without a Jastrow factor. However, the rate of convergence would be exceedingly slow, because the Slater determinants lack singularities present in the wavefunction, namely, the cusps at electron-electron coincidence points. These cusps can already be built into the wavefunction at the level of the two-body correlations by including a Jastrow factor. Incorporating three-body correlations is clearly very advantageous, but the results show that inclusion of four-body correlations may not be the most economical next step to further improvement of the wavefunctions, at least for the heavier systems. Instead, it may be preferable to include more determinants in the wavefunction. Once a sufficiently large number of determinants have been included, it seems likely that it may again become more economical to improve the Jastrow part by including the four-body correlations. Such explorations are needed in order to fully exploit the flexibility that quantum Monte Carlo offers over conventional quantum chemistry methods for the construction of accurate, yet relatively compact, wavefunctions.

5. Acknowledgements

I thank Peter Nightingale for numerous very valuable discussions. The work, reviewed in this chapter, on multiconfiguration wavefunctions and four-body correlations was mostly performed by Claudia Filippi and Chien-Jung Huang, respectively,

A. Appendix

In this appendix we describe the explicit functional forms used for the functions U_i, V_i and W_i of Eq. (25), used in the spherical polar coordinate directed Metropolis method. For each electron, the angles θ, ϕ are measured in coordinates such that the origin is at the nearest nucleus, the z axis is along the direction from the nearest nucleus to the initial position of the electron under consideration and

$$\mathbf{F}_k = \frac{\nabla_\mathbf{k} \Psi(\mathbf{R}_\mathrm{f})|_{\mathbf{R}_\mathrm{f}=\mathbf{R}_\mathrm{i}}}{\Psi(\mathbf{R}_\mathrm{i})} \tag{40}$$

is in the $x-z$ plane with the $x-$component being positive. Hence $\theta_{k,\mathrm{i}} = 0$.

A.1. CHOICE OF $U(r_{k,\mathrm{f}}|\mathbf{R}_\mathrm{i})$

$$U(r_{k,\mathrm{f}}|\mathbf{R}_\mathrm{i}) = \begin{cases} (1+ar_{k,\mathrm{f}})\exp(-\zeta r_{k,\mathrm{f}}), & \text{if } r_{k,\mathrm{i}}/\Delta_r \le r_{k,\mathrm{f}} \le r_{k,\mathrm{i}}\,\Delta_r \\ 0, & \text{otherwise} \end{cases}$$

This form is chosen because, (i) for appropriate choices of a and ζ, it is flexible enough to mimic the behavior of Ψ both near and far from nodes, (ii) the integral of $\sqrt{r_{k,\mathrm{f}}}\,U(r_{k,\mathrm{f}}|\mathbf{R}_\mathrm{i})$ is known in terms of incomplete gamma functions and (iii) its maximum value in the interval $r_{k,\mathrm{i}}/\Delta_r \le r_{k,\mathrm{f}} \le r_{k,\mathrm{i}}\,\Delta_r$ (needed for sampling by rejection) can be easily obtained. For $a = 0$ it reduces to an exponential and for $\zeta = 0$ it reduces to a linear function. The parameters a and ζ are determined by solving the quadratic equation for ζ resulting from matching the radial log-derivatives of $U(r_{k,\mathrm{f}}|\mathbf{R}_\mathrm{i})$ at the initial point and the nearest nucleus to those of the trial wavefunction,

$$\frac{\left.\frac{dU(r_{k,\mathrm{f}}|\mathbf{R}_\mathrm{i})}{dr_{k,\mathrm{f}}}\right|_{r_{k,\mathrm{f}}=r_{k,\mathrm{i}}}}{U(r_{k,\mathrm{i}}|\mathbf{R}_\mathrm{i})} = F_{r,k} \tag{41}$$

$$\frac{\left.\frac{dU(r_{k,\mathrm{f}}|\mathbf{R}_\mathrm{i})}{dr_{k,\mathrm{f}}}\right|_{r_{k,\mathrm{f}}=0}}{U(0|\mathbf{R}_\mathrm{i})} = -Z. \tag{42}$$

where $F_{r,k} = \hat{r}_{k,\mathrm{i}} \cdot \mathbf{F}_k$. the radial log-derivative at the initial point. If the roots of the quadratic equation are real, the smallest positive root is chosen. If both roots are negative or complex, then, if $F_{r,k}$ is negative, ζ is set equal $-F_{r,k}$, while, if $F_{r,k}$ is positive, ζ is set equal to unity. In either case a is determined by matching the radial log-derivative at the initial point. In order to sample$\sqrt{r_{k,\mathrm{f}}}\;|U(r_{k,\mathrm{f}}|\mathbf{R}_\mathrm{i})|$ by rejection it is necessary to determine the its maximum value in the interval $r_{k,\mathrm{i}}/\Delta_r \le r_{k,\mathrm{f}} \le r_{k,\mathrm{i}}\,\Delta_r$, which also requires solving a quadratic equation and checking whether one of the roots or one of the end points yields the maximum value.

A.2. CHOICE OF $V(\theta_{k,\mathrm{f}}|r_{k,\mathrm{f}}, \mathbf{R}_\mathrm{i})$

$V(\theta_{k,\mathrm{f}}|r_{k,\mathrm{f}}, \mathbf{R}_\mathrm{i})$ is chosen to be

$$V(\theta_{k,\mathrm{f}}|r_{k,\mathrm{f}}, \mathbf{R}_\mathrm{i}) = \begin{cases} 1, & \text{if } \theta_{k,\mathrm{f}} \le \theta_\mathrm{M} \\ 0, & \text{otherwise} \end{cases}$$

where θ_M is defined in Eq. (26).

A.3. CHOICE OF $W(\phi_{k,\mathrm{f}}|\theta_{k,\mathrm{f}}, r_{k,\mathrm{f}}, \mathbf{R}_\mathrm{i})$

To match the angular component of the log-derivative, $W(\phi_{k,\mathrm{f}}|\theta_{k,\mathrm{f}}, r_{k,\mathrm{f}}, \mathbf{R}_\mathrm{i})$ is chosen to be

$$W(\phi_{k,\mathrm{f}}|\theta_{k,\mathrm{f}}, r_{k,\mathrm{f}}, \mathbf{R}_\mathrm{i}) = 1 + \min\{F_{\phi,k}\ , 1\}\ \cos\phi_{k,\mathrm{f}}$$

where $F_{\phi,k} = F_{x,k}\, r_{\mathrm{av}} \sin\theta_{k,\mathrm{f}}$ and $F_{x,k} = |\mathbf{F}_k - F_{r,k}\hat{r}_k|$. With these choices of $U(r_{k,\mathrm{f}}|\mathbf{R}_\mathrm{i})$, $V(\theta_{k,\mathrm{f}}|r_{k,\mathrm{f}}, \mathbf{R}_\mathrm{i})$ and $W(\phi_{k,\mathrm{f}}|\theta_{k,\mathrm{f}}, r_{k,\mathrm{f}}, \mathbf{R}_\mathrm{i})$, the log-derivatives of $\Phi(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ match those of $\Psi(\mathbf{R}_\mathrm{f})$ except when $F_{\phi,k} > 1$.

A.4. NORMALISATION

From Eq. (25) it follows that it suffices to calculate

$$I(\mathbf{R}_\mathrm{f}, \mathbf{R}_\mathrm{i}) = \prod_{k=1}^{n}\ I_{\phi,k}\ I_{\theta,k}(r_{k,\mathrm{f}}, \mathbf{R}_\mathrm{i})\ I_{r,k}(\mathbf{R}_\mathrm{i})$$

where

$$\begin{aligned} I_{r,k}(\mathbf{R}_\mathrm{i}) &= \int_{r_{k,\mathrm{i}}/\Delta_r}^{r_{k,\mathrm{i}}\Delta_r} \mathrm{d}r_{k,\mathrm{f}}\ \sqrt{r_{k,\mathrm{f}}}\ |U(r_{k,\mathrm{f}}|\mathbf{R}_\mathrm{i})| \\ &= \begin{cases} \left|G(\tfrac{3}{2}, \zeta r_{k,\mathrm{i}}\Delta_r, -\tfrac{\zeta}{a})\right| + \left|G(\tfrac{3}{2}, -\tfrac{\zeta}{a}, \tfrac{\zeta r_{k,\mathrm{i}}}{\Delta_r})\right| & ,\ \tfrac{r_{k,\mathrm{i}}}{\Delta_r} < -a < r_{k,\mathrm{i}}\ \Delta_r \\ \left|G(\tfrac{3}{2}, \zeta r_{k,\mathrm{i}}\Delta_r, \tfrac{\zeta r_{k,\mathrm{i}}}{\Delta_r})\right| & ,\ \text{otherwise} \end{cases} \end{aligned}$$

$$G(n, x, y) = \frac{\gamma(n, x) - \gamma(n, y)}{\zeta^n} + a\frac{\gamma(n+1, x) - \gamma(n+1, y)}{\zeta^{n+1}}\ ,$$

γ is the incomplete gamma function, and

$$\begin{aligned} I_{\theta,k}(r_{k,\mathrm{f}}, \mathbf{R}_\mathrm{i}) &= \int_0^{\theta_\mathrm{M}} \mathrm{d}\theta_{k,\mathrm{f}}\ \sin\theta_{k,\mathrm{f}} = 1 - \cos\theta_\mathrm{M} \\ I_{\phi,k} &= \int_0^{2\pi} \mathrm{d}\phi_{k,\mathrm{f}}\ (1 + \min\{F_{\phi,k}\ , 1\}\cos\phi_{k,\mathrm{f}}) = 2\pi\ . \end{aligned}$$

For atoms $I_{\theta,k}(r_{k,\mathrm{f}}, \mathbf{R}_\mathrm{i}) = I_{\theta,k}(r_{k,\mathrm{i}}, \mathbf{R}_\mathrm{f})$ because $r_{k,\mathrm{i}}$ and $r_{k,\mathrm{f}}$ appear symmetrically in θ_M [see Eq. (26)]. Hence the factors of $I_{\theta,k}(r_{k,\mathrm{f}}, \mathbf{R}_\mathrm{i})$ and $I_{\theta,k}(r_{k,\mathrm{i}}, \mathbf{R}_\mathrm{f})$ may be omitted in Eq. (15) but for molecules and solids they must be included. If, instead of choosing θ_M according Eq. (26), it is chosen to be a constant then $I(\mathbf{R}_\mathrm{f}, \mathbf{R}_\mathrm{i})$ reduces to $I(\mathbf{R}_\mathrm{i})$ of Eq. (13), but in any case $I(\mathbf{R}_\mathrm{f}, \mathbf{R}_\mathrm{i})$ is the correct normalization.

References

1. W. L. McMillan, Phys. Rev. 138, A442 (1965).

2. D. Ceperley, G.V. Chester and M.H. Kalos, *Phys. Rev. B* **16**, 3081 (1977).
3. There exist methods employing quasi-random numbers, which are more uniformly distributed than random numbers, that converge faster than $1/\sqrt{N}$. See *e.g.* article by Mascagni in this volume and references therein. However, they typically are useful only in an intermediate number of dimensions.
4. N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.M. Teller and E. Teller, *J. Chem. Phys.* **21**, 1087 (1953).
5. We use *probability distribution* to mean either probability, when dealing with a discrete system or, probability density, when dealing with a continuous system. This differs from the terminology sometimes used that a distribution is a cumulative probability density. We will use discrete and continuous notation interchangeably and it will be understood that in most of the expressions, matrices may be replaced by integral kernels, probabilities by probability densities, and summations by integrations and vice versa.
6. M.H. Kalos and P.A. Whitlock, *Monte Carlo Methods,* Vol. 1, (Wiley, 1986).
7. C.J. Everett and E.D. Cashwell, *A third Monte Carlo Sampler*, Los Alamos Technical Report No., LA-9721-MS, 1983.
8. For example, in an electronic structure calculation of an atom, states with an electron at infinity cannot be, and need not be, accessed in in a finite number of Monte Carlo steps.
9. A matrix is cyclic if the state space can be divided into subspaces, such that there is a non-zero probability only for transitions from states in one subspace to states of the next subspace.
10. C.J. Umrigar, *Phys. Rev. Lett.* **71**, 408 (1993).
11. W.K. Hastings, *Biometrika* **57**, 97 (1970).
12. M. Rao and B.J. Berne, *J. Chem. Phys.* **71**, 129 (1979), introduced an adjustable parameter λ in their force-bias method. Although they do not describe their work in these terms, note that they obtained better results for $\lambda = 1/2$ corresponding to $S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) \overset{\propto}{\sim} \sqrt{\rho(\mathbf{R}_\mathrm{f})}$ than for $\lambda = 1$ corresponding to $S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) \overset{\propto}{\sim} \rho(\mathbf{R}_\mathrm{f})$.
13. C.J. Umrigar, unpublished.
14. The terms *valence electron* or *core electron* are only used to indicate the instantaneous distance of an electron from the nearest nucleus. During the course of a Monte Carlo run a given electron will be in both the valence and core regions part of the time.
15. The idea of minimizing the variance of the local energy goes back to at least 1935. J.H. Bartlett, J.J. Gibbons and C.G. Dunn: *Phys. Rev.* **47**, 679 (1935) credit it to J. Boussinesq: *Theorie de la chaleur* (I). In the context of Monte Carlo calculations, the method was first used by R. L. Coldwell, *Int. J. Quantum. Chem. Symp.* **11**, 215 (1977).
16. C.J. Umrigar, K.G. Wilson and J.W. Wilkins, *Phys. Rev. Lett.* **60**, 1719 (1988); in *Computer Simulation Studies in Condensed Matter Physics, Recent Developments,* ed. by D.P. Landau K.K. Mon and H.B. Schüttler, Springer Proc. Phys. (Springer, Berlin, 1988).
17. C.J. Myers, C.J. Umrigar, J.P. Sethna and J.D. Morgan, *Phys. Rev. A* **44**, 5537 (1991).
18. M.P. Nightingale and C.J. Umrigar, unpublished. The program is available upon request.
19. M. Meierovich, A. Mushinski, and M.P. Nightingale, *J. Chem. Phys.* **105**, 6498 (1996), and URL http://xxx.lanl.gov/abs/chem-ph/9512001.
20. M. P. Nightingale and H.W.J. Blöte, *Phys. Rev. Lett.* **80**, 1007 (1998).
21. M.P. Nightingale and H.W.J. Blöte, *Phys. Rev. B* **54**, 1001 (1996); URL http://xxx.lanl.gov/abs/cond-mat/9602089.
22. M.P. Nightingale and H.W.J. Blöte, *Phys. Rev. Lett.* **76**, 4548 (1996); URL http://xxx.lanl.gov/abs/cond-mat/9601059.

23. Chien-Jung Huang, Claudia Filippi, and C. J. Umrigar, *J. Chem. Phys.* **108**, 8838 (1998).
24. T. Kato, Commun. Pure Appl. Math. **10**, 151 (1957).
25. R.T. Pack and W. Byers-Brown, J. Chem. Phys. **45**, 556 (1966).
26. J.D. Morgan, III, and W. Kutzelnigg, J. Chem. Phys. **97**, 2425 (1993).
27. Claudia Filippi and C.J. Umrigar, J. Chem. Phys. **105**, 213 (1996).
28. A. Mushinski and M.P. Nightingale, J. Chem. Phys. **101**, 8831 (1994).
29. Chien-Jung Huang, C.J. Umrigar and M.P. Nightingale, J. Chem. Phys. **107**, 3007 (1997).
30. E.R. Davidson, S.A. Hagstrom, S.J. Chakravorty, V.M. Umar and C.F. Fischer, Phys. Rev. A **44**, 7071 (1991); S.J. Chakravorty, S.R. Gwaltney, E.R. Davidson, F.A. Parpia and C.F. Fischer, Phys. Rev. A **47**, 3649 (1993).