1-1-2006

Population size bias in descendant-weighted diffusion quantum Monte Carlo simulations

G. Lee Warren
University of Tennessee, Knoxville

Robert J. Hinde
University of Tennessee, rhinde@utk.edu

Population size bias in descendant-weighted diffusion quantum Monte Carlo simulations

G. Lee Warren and Robert J. Hinde*
Department of Chemistry, University of Tennessee, Knoxville, Tennessee 37996, USA
(Received 28 July 2005; published 22 May 2006)

We consider the influence of population size on the accuracy of diffusion quantum Monte Carlo simulations that employ descendant weighting or forward walking techniques to compute expectation values of observables that do not commute with the Hamiltonian. We show that for a simple model system, the d-dimensional isotropic harmonic oscillator, the population size must increase rapidly with d in order to ensure that the simulations produce accurate results. When the population size is too small, expectation values computed using descendant-weighted diffusion quantum Monte Carlo simulations exhibit significant systematic biases.

DOI: 10.1103/PhysRevE.73.056706 PACS number(s): 02.70.Ss

I. INTRODUCTION

The diffusion quantum Monte Carlo (DQMC) method [1] is a stochastic approach for solving the many-body Schrödinger equation by exploiting the isomorphism between this equation and the imaginary-time diffusion equation. In essence, a DQMC simulation follows the time evolution of a set of particles, often called "psips" or "walkers," whose motion through configuration space is governed by a diffusion equation with source and sink terms that are related to the many-body potential energy surface that appears in the corresponding Schrödinger equation. As the simulation progresses in time, transients associated with the initial distribution of particles die out and the long-time time-averaged concentration field of walkers approaches a steady state. This steady state distribution is (in the absence of importance sampling) proportional to the nodeless wave function of lowest energy that solves the Schrödinger equation under consideration.

Any practical DQMC simulation necessarily follows the evolution of a finite collection of walkers, and the steady state distribution of walkers which is sought must be constructed by averaging together many "snapshots" of the walkers' instantaneous distribution. The resulting average only approximately represents the true ground state wave function of the associated many-body quantum system, both because the population of walkers is finite and because the number of snapshots that can be collected in any practical DQMC simulation is also finite. It is natural to expect that this approximation to the true steady state distribution will approach the true distribution more closely as either the number of walkers or the number of snapshots increases. In this paper, we examine the relationship between this convergence process and the dimensionality of the underlying configuration space.

A simple thought experiment serves to motivate our inquiry. Suppose that we select N random numbers $\{x_i\}$ from a reasonably well behaved distribution function $F(x)$. We then show the set $\{x_i\}$ and the function $F(x)$ to an impartial referee and ask the referee whether the set of numbers is consistent with the specified distribution. Next, we select N pairs $\{(x_i,y_i)\}$ from a bivariate distribution $G(x,y)$. Again, we show the set of points and the distribution function $G(x,y)$ to the referee and ask whether the set of points is consistent with the specified distribution. We continue this game, each time selecting N points from a multivariate distribution whose dimension increases by one in each round of the game. It is not difficult to suppose that once the dimensionality of the space is sufficiently high, the referee will quit the game in frustration because the collection of N points does not cover the multidimensional space sufficiently densely to provide any real sense of the underlying distribution function. At that point, if we want to continue the game, we must increase the number of points N that we select from the distribution function. The game we have described is one that we ourselves play whenever we try to interpret the approximate steady state distribution obtained by averaging snapshots collected during a DQMC simulation.

Here we investigate this phenomenon using a simple model system: the d-dimensional isotropic harmonic oscillator. We begin in Sec. II by describing a method for quantifying the similarity between this system's ground state wave function and a snapshot collected from a DQMC simulation with N walkers. We demonstrate that with this definition of the similarity, N must grow exponentially with d in order to guarantee that an individual snapshot closely resembles the underlying wave function. Section III then shows that one consequence of this is that, unless very large walker populations are used, DQMC simulations that employ descendant weighting or forward walking techniques to estimate the expectation values of coordinate-space observables that do not commute with the Hamiltonian can give substantially biased results when d is large [2]. In Sec. IV we demonstrate that the use of importance sampling techniques [3] can reduce, but not eliminate, these biases. We conclude with a brief discussion in Sec. V.

II. IS A SINGLE SNAPSHOT SIMILAR TO THE WAVE FUNCTION?

We consider a set of d independent, identical one-dimensional harmonic oscillators. The oscillators' configuration is specified by the vector $\mathbf{q}=(q_1,q_2,\ldots,q_d)$ of dimen-

*Electronic address: rhinde@utk.edu

sionless coordinates, and the potential energy of the set of oscillators is given by $V(\mathbf{q})=\hbar\omega q^2/2$. Henceforth we employ atomic units and set $\omega=1$ a.u.

We suppose that we have a DQMC simulation algorithm that generates snapshots of the locations of N walkers; a single snapshot is simply a collection $\{\mathbf{q}_i: i=1,2,\ldots,N\}$ of N positions in $\mathbb{R}^d$. We further suppose that this algorithm is accurate in the sense that the walker locations in any particular snapshot are drawn from the normalized distribution $F(\mathbf{q})=(2\pi)^{-d/2}\exp(-q^2/2)$, which is proportional to the system's ground state wave function $\Psi(\mathbf{q})$. Let

$$S(\mathbf{q}) = 2^{d/2}\exp(-q^2/2) \tag{1}$$

so that

$$\int_{\mathbb{R}^d} S(\mathbf{q})F(\mathbf{q})d\mathbf{q} = 1. \tag{2}$$

For a snapshot $\{\mathbf{q}_i: i=1,2,\ldots,N\}$ of N walkers drawn from the distribution $F(\mathbf{q})$, we define

$$s = \frac{1}{N}\sum_{i=1}^{N} S(\mathbf{q}_i), \tag{3}$$

which is a Monte Carlo approximation to the left-hand side of Eq. (2), to be the *similarity* between the snapshot and the distribution $F(\mathbf{q})$. We expect that snapshots that faithfully resemble the distribution $F(\mathbf{q})$ will have s values near 1.

For single-walker snapshots ($N=1$), the observed values of s cover the range $0\leqslant s\leqslant s_{\max}$, with $s_{\max}=2^{d/2}$, and follow the probability distribution

$$P(s) = \frac{[\ln(s_{\max}/s)]^{d/2-1}}{s_{\max}\Gamma(d/2)}. \tag{4}$$

This distribution has unit mean for all values of d and a variance $\sigma^2=(2/\sqrt{3})^d-1$ that grows exponentially with d. For $N>1$, the distribution's variance σ^2/N also grows exponentially with d, as shown in Fig. 1. Thus for a snapshot of N walkers to have a high probability of resembling (in the sense quantified by s) the true steady state distribution $F(\mathbf{q})$, N must grow exponentially with d.

III. COMPUTING OBSERVABLES FROM DQMC SNAPSHOTS

What implications does this finding have for practical DQMC simulations? To answer this question, we turn to the problem of using the DQMC method to evaluate the expectation value $\langle B\rangle=\langle\Psi(\mathbf{q})|B(\mathbf{q})|\Psi(\mathbf{q})\rangle$ of a nondifferential operator $B(\mathbf{q})$ that does not commute with the Hamiltonian. (Henceforth we call such operators Hamiltonian-noncommuting, or HNC, operators.) This expectation value can be estimated from an individual DQMC snapshot as a weighted average of the observable's value over the walkers comprising the snapshot:

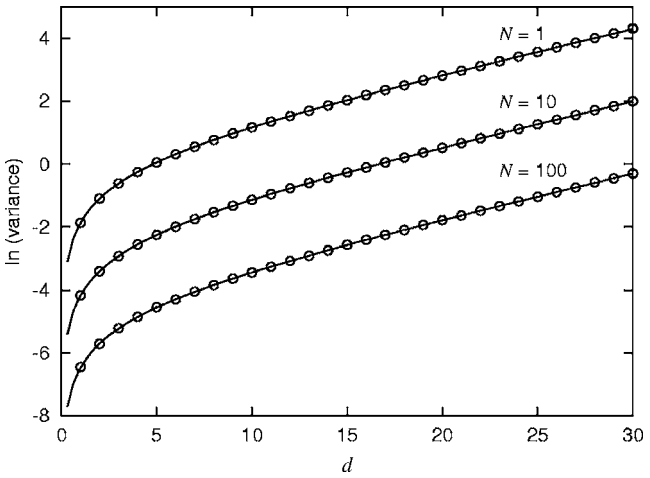

FIG. 1. Semilog plot of the variance of the similarity distribution $P(s)$ as a function of dimension d and number of walkers N. Each point is computed from an ensemble of $5\times10^8/N$ randomly sampled N-walker snapshots. The function $\ln[(2/\sqrt{3})^d-1)/N]$ is plotted for three values of N using solid lines.

$$B = \left(\sum_{i=1}^{N} B(\mathbf{q}_i)W(\mathbf{q}_i)\right)\left(\sum_{i=1}^{N} W(\mathbf{q}_i)\right)^{-1}. \tag{5}$$

We have dropped the angle brackets to indicate that B is an approximation to, and distinct from, the expectation value $\langle B\rangle$. The weight factor $W(\mathbf{q})$ in Eq. (5) is, in the absence of importance sampling, a quantity proportional to the value $\Psi(\mathbf{q})$ of the wave function at the position $\mathbf{q}$; this quantity can be estimated during a DQMC simulation using forward walking [4] or descendant weighting [5] techniques. (We discuss importance-sampled simulations below in Sec. IV.)

Consider an idealized DQMC simulation in which this weight factor can be computed exactly for each walker in the snapshot. Returning to our model system consisting of d uncoupled harmonic oscillators, we can, without loss of generality, set $W(\mathbf{q})=S(\mathbf{q})$ as given in Eq. (1). With this choice of $W(\mathbf{q})$, Eq. (5) can be written as $B=b/s$, where

$$b = \frac{1}{N}\sum_{i=1}^{N} B(\mathbf{q}_i)S(\mathbf{q}_i). \tag{6}$$

We now ask what shape the distribution $P(B)$ of observed B values has, and examine how this shape depends on the number of points N in the DQMC snapshot.

To answer these questions, we first consider the distribution $P(b)$ of b values generated by the sum in Eq. (6). We examine the specific HNC operator $B(\mathbf{q})=q^2/d$, which has expectation value $\langle B\rangle=1/2$. The distribution $P(b)$ has mean $1/2$ and variance $[(2/\sqrt{3})^d(2+d)/9d-1/4]/N$ that grows exponentially with d. We therefore conclude that for a snapshot with N walkers to produce a value for b that falls close to the mean value of $1/2$, N must grow exponentially with d.

We therefore anticipate that for small N and large d, the distribution $P(B)$ of observed values of B will be rather broad, because (1) both the numerator b and denominator s of B have variances that increase exponentially with d and

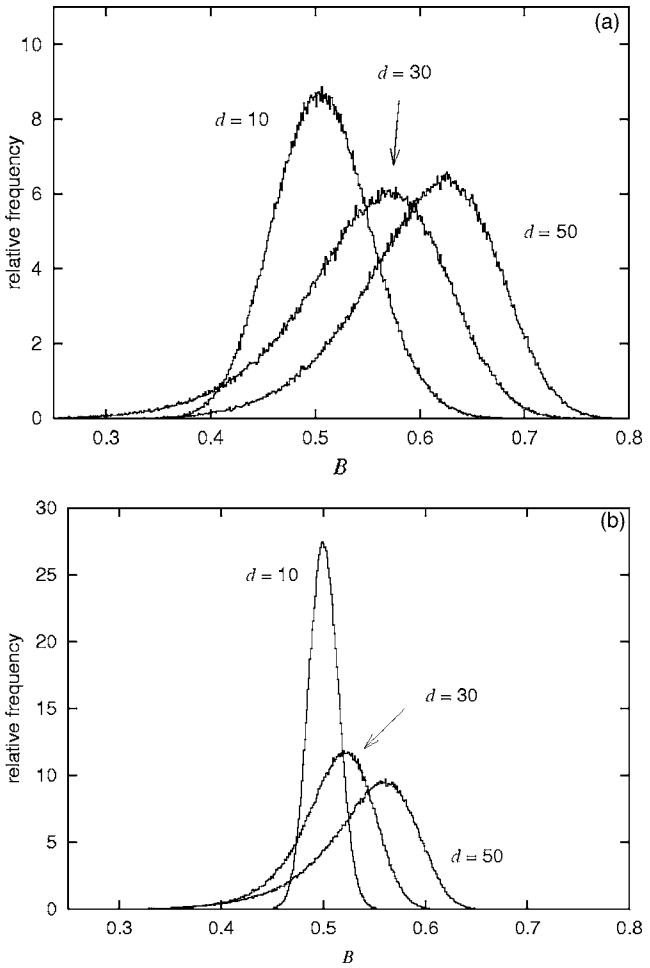

FIG. 2. Probability distribution function $P(B)$ for the quantity $B=q^2/d$ as a function of dimension d and number of walkers N. Panel (a) shows $P(B)$ for $N=100$; panel (b) shows $P(B)$ for $N=1000$. Each distribution function is normalized to unit area and is computed from an ensemble of 5×10^5 randomly sampled N-walker snapshots. Note that the horizontal axis is identical in both panels.

(2) the b and s values for typical snapshots are only weakly correlated with each other. Evidence for this proposition is shown in Fig. 2, which shows how the distribution function $P(B)$ depends on d at both $N=100$ [panel (a)] and $N=1000$ [panel (b)]. At a fixed value of d, $P(B)$ becomes more sharply peaked as N increases, while at a fixed value of N, $P(B)$ broadens as d increases.

Figure 2 also indicates that for large d, the mean of the distribution $P(B)$ deviates substantially from the true expectation value $\langle B\rangle=1/2$. Consequently, for this model system, the approximate expectation value computed via descendant-weighted DQMC will exhibit a systematic error when d is large. Figure 3 shows how this systematic error depends on the number of walkers N; although the mean of $P(B)$ approaches the true expectation value of $1/2$ as N increases, the rate of convergence of this mean with N is rather slow for large values of d.

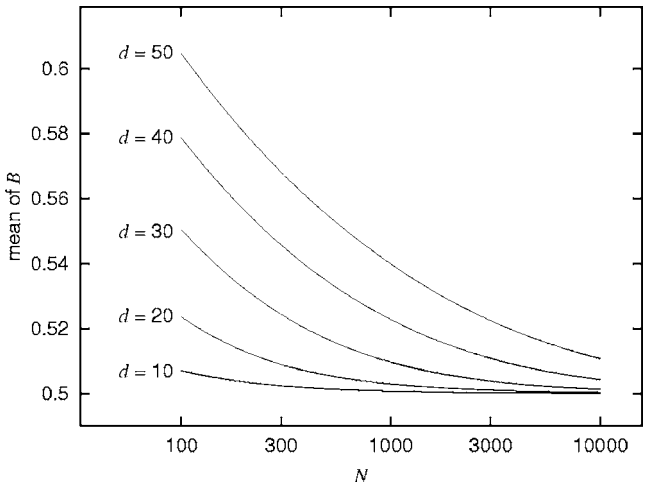

FIG. 3. Mean value of B as a function of N for five values of d. For a given N and d, this mean is computed from an ensemble of 5×10^5 randomly sampled N-walker snapshots. Note that the horizontal axis is logarithmic.

IV. THE ROLE OF IMPORTANCE SAMPLING

In importance-sampled DQMC simulations, a trial function $\Psi_{\rm tr}(\mathbf{q})$ that resembles the ground state wave function is used to guide the walkers' time evolution [3]. The long-time steady state distribution of the walkers is then proportional to the product $\Psi(\mathbf{q})\Psi_{\rm tr}(\mathbf{q})$ of the ground state wave function and the trial function, while the weight function $W(\mathbf{q})$ that appears in Eq. (5) is proportional to the ratio $\Psi(\mathbf{q})/\Psi_{\rm tr}(\mathbf{q})$. Here we show that a good trial function can temper somewhat the population size requirements described in the preceding two sections; however, for our model system, these population size requirements seem to vanish only in the limit of perfect importance sampling, when $\Psi_{\rm tr}(\mathbf{q})=\Psi(\mathbf{q})$.

We take the trial function to be $\Psi_{\rm tr}(\mathbf{q})=\exp(-\alpha q^2/2)$ where the parameter α controls the quality of the trial function. For $\alpha=0$ we have no importance sampling ($\Psi_{\rm tr}=1$), while for $\alpha=1$ we have perfect importance sampling. With this trial function, the long-term steady state distribution of walkers is given by

$$F_\alpha(\mathbf{q})=\left(\frac{1+\alpha}{2\pi}\right)^{d/2}\exp[-(1+\alpha)q^2/2] \tag{7}$$

and the weight factor that appears in Eq. (5) becomes

$$W_\alpha(\mathbf{q})=\left(\frac{2}{1+\alpha}\right)^{d/2}\exp[-(1-\alpha)q^2/2]. \tag{8}$$

Following our earlier discussion, the approximate expectation value of the operator $B(\mathbf{q})$, estimated using importance-sampled DQMC simulations, can be written as $B_\alpha=b_\alpha/s_\alpha$, where

$$b_\alpha=\frac{1}{N}\sum_{i=1}^{N}B(\mathbf{q}_i)W_\alpha(\mathbf{q}_i) \text{ and } s_\alpha=\frac{1}{N}\sum_{i=1}^{N}W_\alpha(\mathbf{q}_i). \tag{9}$$

For single-walker snapshots ($N=1$), the variances of these quantities can be computed analytically:

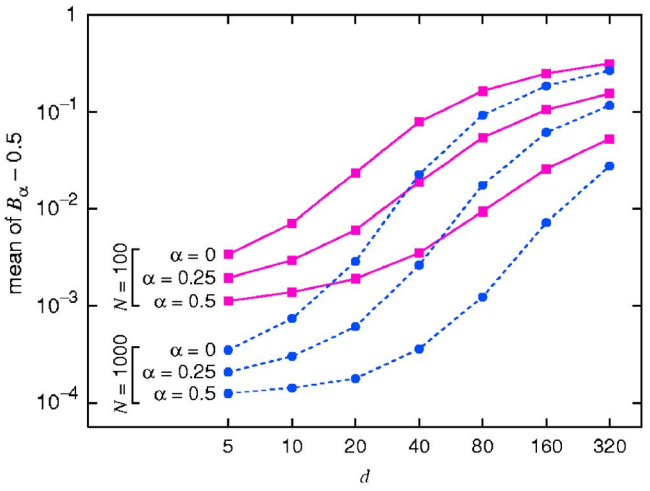

FIG. 4. (Color online) Error in the mean value of B_α as a function of d for two values of N and three values of α. Solid lines and filled squares are for $N=100$; dashed lines and filled circles are for $N=1000$. For a given N, d, and α, the mean value of B_α is computed from an ensemble of 5×10^5 randomly sampled N-walker snapshots. Note that both axes are logarithmic in this figure. Lines are drawn to guide the eye but have no mathematical significance.

$$b_\alpha: \quad \sigma^2 = \gamma^d\left(\frac{2+d}{d(3-\alpha)^2}\right) - \frac{1}{4}, \tag{10}$$

$$s_\alpha: \quad \sigma^2 = \gamma^d - 1, \tag{11}$$

where $\gamma=2/\sqrt{(1+\alpha)(3-\alpha)}$. Except in the limit of perfect importance sampling ($\alpha=1$), the quantity $\gamma>1$ and both variances thus grow exponentially with d once d is large enough. We therefore anticipate that once this critical d value is reached, B_α will exhibit behavior much like that shown for B in Figs. 2 and 3. In particular, we anticipate that for small N and large d, the probability distribution $P(B_\alpha)$ for B_α will be broad and skewed, with a mean that deviates systematically from the exact quantum mechanical expectation value $\langle B\rangle=1/2$.

Figure 4 confirms this hypothesis, and shows how the bias in the mean value of B_α depends on N, d, and α. Although improving the quality of the trial function $\Psi_{\rm tr}(\mathbf{q})$ by increasing α does reduce the bias in B_α at a given N and d, we see that this bias nevertheless increases (although not exponentially) with d for all of the mock DQMC simulations performed here. It thus appears that the requirement that the DQMC population size N must grow rapidly with d is fairly general, and applies to both importance-sampled and nonimportance-sampled DQMC simulations.

V. DISCUSSION

DQMC simulations are generally thought to be particularly useful for finding the ground state of many-body quantum systems, because the computational effort required to simulate the diffusion of walkers in configuration space does not grow exponentially with the dimension of this space. We have shown that this is only part of the story. Unless perfect importance sampling techniques are employed (in which case the exact ground state wave function is known and there is no need to perform a DQMC simulation), it appears that the population of walkers in a DQMC simulation must grow exponentially with the dimension of the underlying configuration space in order for a walker snapshot to faithfully represent the system's ground state wave function. Consequently, it is likely that large walker populations will be needed to compute accurate expectation values of HNC operators for many-body systems using descendant weighting techniques. As an example, we have considered a simple model system, the d-dimensional isotropic harmonic oscillator. For this system, when d is large and the walker population is too small, the expectation values of HNC operators computed via descendent weighting can exhibit large systematic biases. Although importance sampling techniques can reduce these biases, we observe that the demands placed on the quality of the trial wave function also increase as d increases.

These findings suggest that caution should be employed when using forward walking or descendant weighting techniques within a DQMC simulation to estimate expectation values of HNC operators in systems with high-dimensional configuration spaces. For these systems, other methods for estimating expectation values of HNC operators, such as variational path integral simulations [6] or reptation quantum Monte Carlo simulations [7], are likely to be more efficient than descendant-weighted DQMC simulations.

ACKNOWLEDGMENTS

This work was supported by the Air Force Office of Scientific Research through Grant No. F-49620-01-1-0068, by the Donors of the Petroleum Research Fund, administered by the American Chemical Society, by the National Science Foundation through Grant No. CHE-0414705, and by the UTK Center for Information Technology Research. G.L.W. also received support from the National Science Foundation.

[1] J. B. Anderson, J. Chem. Phys. **63**, 1499 (1975).

[2] These systematic biases are different from those discussed by Cerf and Martin [N. Cerf and O. C. Martin, Phys. Rev. E **51**, 3679 (1995)], which are associated with population *control* mechanisms in DQMC simulations. Cerf and Martin showed that employing population control makes the stationary distribution of a DQMC simulation a biased estimate of a system's true ground state wave function; we show here that expectation values suffer from a systematic bias *even when* the collection of walkers is drawn from the true ground state wave function.

[3] M. H. Kalos, D. Levesque, and L. Verlet, Phys. Rev. A **9**, 2178 (1974).
[4] R. N. Barnett, P. R. Reynolds, and W. A. Lester, Jr., J. Comput. Phys. **96**, 258 (1991).
[5] J. Casulleras and J. Boronat, Phys. Rev. B **52**, 3654 (1995).
[6] D. M. Ceperley, Rev. Mod. Phys. **67**, 279 (1995).
[7] S. Baroni and S. Moroni, Phys. Rev. Lett. **82**, 4745 (1999).