JOURNAL OF COMPUTATIONAL PHYSICS 74, 127–142 (1988)

Statistical Error of Diffusion Monte Carlo

STUART M. ROTHSTEIN AND JAN VRBIK

Departments of Chemistry and Mathematics,
Brock University, St. Catharines, Ontario, Canada L2S 3A1

Received December 8, 1984; revised February 13, 1987

We show that the standard error of the Diffusion Monte Carlo estimators varies (for small values of τ) with $\tau^{-1/2}$, where τ is the value of the time step. This has serious implication for the design and optimization of specific simulations. The statistical variation is analyzed in terms of three components which relate to guiding function ψ_T, branching, and serial correlation of consecurive iterations. The behavior of these is investigated empirically as a function of τ. An appropriate scheme (applicable to any simulation calculation with serially correlated data) is proposed to facilitate reliable estimation of the statistical error. © 1988 Academic Press, Inc.

This paper concerns the diffusion Monte Carlo (DMC) method [1–4] for obtaining an estimate of the exact ground-state energy of a molecule. Several calculations have been reported with progressively improved accuracy. However, the results for some molecules exhibit a relatively large statistical error; it has been estimated that for LiH results to be competitive with current *ab initio* methods could require a 10-fold reduction of the error [5]. The long term objective of our continuing research is to design a vastly more efficient and competitive procedure.

In this paper we analyze the behavior of the statistical error as a function of time step τ. Although it is known (by inappropriately assuming a Markovian series of observations) that the variance of a serially correlated variable in Metropolis sampling varies as τ^{-1} [6] suggesting a similar result for DMC sampling, a more general and detailed analysis has been lacking. The results of our analysis have serious implications for the design of efficient simulations as well as for isolating the essential ingredients for efficient guiding functions. (Optimization of DMC is a topic which we consider in another publication [7].) We also address the problem of efficiently estimating the statistical error, without recourse to the commonly used, yet inefficient technique of creating huge blocks of data which may be treated as statistically independent observations.

We assume familiarity with the DMC technique for finding the ground-state eigenvalue E_0 of the Schrödinger equation (expressed in atomic units)

$$-\tfrac{1}{2}\nabla^2\phi_0 + V(\mathbf{R})\phi_0 = E_0\phi_0 \tag{1}$$

by converting it to

$$[-\tfrac{1}{2}\nabla^2 + \nabla\cdot\mathbf{F}(\mathbf{R}) + E_L(\mathbf{R})]\, f(\mathbf{R}) = E_0 f(\mathbf{R}), \tag{2}$$

0021-9991/88 $3.00

581/74/1-9

where ∇ operates on both functions to its right. The two functions

$$\mathbf{F}(\mathbf{R}) = \psi_T^{-1} \nabla \psi_T \tag{3}$$

and

$$E_L = -\tfrac{1}{2}(\nabla \cdot \mathbf{F}(\mathbf{R}) + F^2) + V(\mathbf{R}) \tag{4}$$

are calculated using a trial solution ψ_T to (1).

There are several variations of the basic procedure, but for the analysis of the leading term of the statistical error they all are equivalent to the following:

(a) Select a value of time step τ (here measured in atomic units) and generate a set of M random $3N$-dimensional configurations $\mathbf{R}$, each of which specifies the coordinates of the N particles of the system under investigation.

(b) Advance the configurations to their new values

$$\mathbf{R}_i^a = \mathbf{R}_i + \tau \mathbf{F}(\mathbf{R}_i) + \tau^{1/2} \chi_i + \cdots, \qquad i = 1, 2, ..., M, \tag{5}$$

where each χ_i is a $3N$-dimensional random vector with independent components drawn from a distribution with zero mean and unit variance. The three dots in (5), and from now on, represent higher order terms which may vary between different versions of DMC.

(c) *Branching*; create a new set of configurations by including

$$\text{int}[1 - \tau(E_L(\mathbf{R}_i) - E_{\text{av}}) + \cdots + u_i] \tag{6}$$

copies of $\mathbf{R}_i^a$, where u_i is a random number uniformly distributed over (0, 1). E_{av} is the corresponding iteration average of the local energy, namely,

$$E_{\text{av}} = \sum_{i=1}^{M} E_L(\mathbf{R}_i)/M. \tag{7}$$

This provides a single estimate of $E(\tau)$, the exact eigenvalue of the problem actually simulated (which equals E_0 in (2) plus a small τ-dependent bias introduced by having to approximate the corresponding Green function [8, 9]). Note that the first two terms in brackets of (6) are, within our accuracy, equivalent to the more recognizable

$$\exp\{-\tau[(E_L(\mathbf{R}_i) + E_L(\mathbf{R}_i^a))/2 - E_{\text{av}}]\}. \tag{8}$$

(d) If the resulting number of new configurations differs from M, adjust it to *exactly* M by uniform random deletion or duplication of the same (a necessary stabilization of the process). All $\mathbf{R}^a$ which survive this step are denoted $\mathbf{R}'$.

(e) Steps (b)–(d) complete one iteration of the simulation procedure. These are repeated K times, and the grand-mean, E_{gm}, of the individual iteration averages E_{av} is computed to be used as the overall estimate of $E(\tau)$.

An ultimately unbiased estimate of E_0 may be obtained by repeating steps (a)–(e), using several distinct values of τ, then extrapolating to $\tau = 0$ by means of statistical regression. This requires a good understanding of $E(\tau)$ as a function of τ (see [7–9]). To minimize the standard error of the regression intercept some knowledge of the behavior of the *variance* of E_{gm} is required as well. To obtain this is a major objective of our paper.

Error Analysis

The statistical error of E_{gm} is due to two major factors: (a) the variance of the individual iteration averages E_{av} and (b) the strong serial correlation of *consecutive* E_{av} values. Both of these must be accurately estimated and properly included in calculating the error bars. To provide a good mathematical description of their τ-dependence is the concern of this section. We will assume that the time step is sufficiently small, so that only the leading term in each τ-expansion of a function of τ need be considered. The range of applicability of this approximation will be illustrated by numerical examples in the following section.

We know that, after sufficiently many iterations, the set of M configurations obtained at the end of each iteration is a random (not independent!) sample from a distribution with the pdf equal to $f(\mathbf{R}; \tau)$, where

$$\mathscr{T} f(\mathbf{R}; \tau) - \mathscr{O}_\tau f(\mathbf{R}; \tau) = 0, \tag{9}$$

$$\mathscr{T} \equiv -\tfrac{1}{2}\nabla^2 + \mathbf{V} \cdot \mathbf{F}(\mathbf{R}) + E_L(\mathbf{R}) - E_0, \tag{10}$$

and $\mathscr{O}_\tau$ is an operator in the first and higher order of τ [8]. This means that, after competing step (b), the set of configurations is distributed according to

$$f(\mathbf{R}) + (\tfrac{1}{2}\tau\nabla^2 - \tau\mathbf{V} \cdot F(\mathbf{R}))f(\mathbf{R}) + \cdots = [1 + \tau(E_L(\mathbf{R}) - E_0)]f(\mathbf{R}) \tag{11}$$

(Here, and from now on, our notation suppresses the parametric τ-dependence in f.)

Step (c) converts these back into a sample from $f(\mathbf{R})$. For the purpose of our analysis, we describe a process which combines steps (c) and (d) (correct only to order τ):

(i) If $E_L(\mathbf{R}_j) < E_0$, the "deficient case," the new configuration, $\mathbf{R}_j^a$, automatically remains a part of the new sample.

(ii) If $E_L(\mathbf{R}_i) > E_0$, the "excessive case," $\mathbf{R}_i^a$ is kept with the probability of $1 - \tau(E_L(\mathbf{R}_i) - E_0)$. With probability of $\tau(E_L(\mathbf{R}_i) - E_0)$ configuration $\mathbf{R}_i^a$ is replaced by a configuration selected from among the "deficient" set, according to discrete probabilities proportional to $E_0 - E_L(\mathbf{R}_j)$, where j runs over the "deficient" configurations. This can be seen as sampling from the following distribution:

$$f^-(\mathbf{R}) = \frac{\max(0, E_0 - E_L(\mathbf{R}))f(\mathbf{R})}{\int_{\mathbf{R}} (E_0 - E_L(\mathbf{R}))f(\mathbf{R})\, d\mathbf{R}}, \tag{12}$$

where the region $\mathbf{R}_-$ consists of all "deficient" values of $\mathbf{R}$ (those for which $E_L(\mathbf{R}) < E_0$).

Now we can compute the serial correlation between $E_L(\mathbf{R}_i)$ and $E_L(\mathbf{R}_i')$, the local energies for the ith configuration before and after steps (b)–(d), of which the serial correlation of the consecutive E_{av} values is an immediate consequence. By definition, the covariance is given by

$$\text{cov}(E_L(\mathbf{R}_i), E_L(\mathbf{R}_i')) \equiv \iint (E_L(\mathbf{R}_i) - E_0)(E_L(\mathbf{R}_i') - E_0)\, g(\mathbf{R}_i, \mathbf{R}_i')\, d\mathbf{R}_i\, d\mathbf{R}_i', \quad (13)$$

where $g(\mathbf{R}_i, \mathbf{R}_i')$ is the joint distribution (in pdf sense) of $\mathbf{R}_i$ and $\mathbf{R}_i'$ (two highly correlated random variables). As shown in detail in Appendix A, the covariance is given by

$$\text{cov}(E_L(\mathbf{R}_i), E_L(\mathbf{R}_i')) = \text{var}(E_L(\mathbf{R}))(1 - a\tau) + \cdots, \quad (14)$$

where a is a constant (depending only on ψ_T and V) defined by Eq. (A6). The serial correlation coefficient between $E_L(\mathbf{R}_i)$ and $E_L(\mathbf{R}_i')$ is thus equal to

$${}^1\rho = 1 - a\tau + \cdots. \quad (15)$$

We may now repeat the above derivation to obtain the serial correlation between $E_L(\mathbf{R}_i)$ and $E_L(\mathbf{R}_i^{(2)})$, the local energy for the ith configuration, *two* iterations later. In the covariance expression (A1) we now have 2τ appearing instead of τ (as two moves have taken place, the probability of being in region $\mathbf{R}_+$ doubles). Also, to within our accuracy, we have a $\tau\mathbf{F}(\mathbf{R}_i)$ per iteration in (A2). Finally, the sum of two independent random variables of type χ is equivalent to a single χ, with the standard deviation of $\sqrt{2}$. This introduces a factor of 2 into $\frac{1}{2}\tau\nabla^2$ of (A2). Thus it is easy to see that two consecutive applications of the arguments of Appendix A will result in replacing τ by 2τ in all formulas of that appendix. By extension, the serial correlation coefficient between $E_L(\mathbf{R}_i)$ and $E_L(\mathbf{R}_i^{(k)})$, where $\mathbf{R}_i^{(k)}$ is the ith configuration, k iterations later, is given by

$${}^k\rho = 1 - k \cdot a\tau + \cdots. \quad (16)$$

Since ${}^k\rho$ must tend to zero with increasing k, it is more convenient to rewrite (16) as

$${}^k\rho = [\exp(-a\tau + \cdots)]^k = ({}^1\rho)^k. \quad (17)$$

This would effectively imply that the sequence of $E_L(\mathbf{R}_i)$ values is Markovian. However, such a assumption may not be totally adequate for all values of k. Thus, in the most general case, one must rewrite (17) using additional terms, as

$${}^k\rho = A_1[\exp(-a_1\tau + \cdots)]^k + A_2[\exp(-a_2\tau + \cdots)]^k + \cdots, \quad (18)$$

where

$$\sum A_i = 1, \tag{19}$$

$$\sum A_i a_i = a, \tag{20}$$

and a is given in (14). Both (17) and (18) imply exponential decrease and the overall convergence of the ${}^k\rho$ series, which in the former case sums to $(a\tau)^{-1} + \cdots$ and in the latter to $\tau^{-1} \sum A_i/a_i + \cdots$.

To simplify some of our arguments we will occasionally assume (17) to be sufficiently accurate (which is quite often of case), but will make sure that our final scheme for estimating the experimental error bars in free of this "Markovian" assumption.

Now, we want to extend these results to the sequence of K iteration averages E_{av} (7). Unfortunately,

$$\text{var}(E_{\text{av}}) \neq \text{var}(E_L(\mathbf{R}))/M \tag{21}$$

due to intercorrelation of some of the $E_L(\mathbf{R}_i)$ values, as branching (step (c)) results in creating duplicate configurations. This implies that (16) does not apply for ρ_{av}; we will now find its analog.

Proceeding parallel to the analysis which lead to (14) and (17), one can show that the actual relationship between the variance of an iteration average E_{av} and the variance of an individual $E_L(\mathbf{R}_i)$ value is

$$\text{var}(E_{\text{av}}) = [\text{var}(E_L(\mathbf{R})) + 2C]/M + \cdots. \tag{22}$$

This result is derived in Appendix B, with the constant C defined in (B9). C itself is independent of τ and arises from branching, which creates strongly correlated pairs of $E_L(\mathbf{R})$ values (a parallel correlation) to serially propagate through the iterations. Thus C, which may be readily estimated, can provide an assessment of the effect of branching on the statistical error (see Fig. 1, below).

Similarly, in Appendix C we show that the covariance of consecutive E_{av} values is given by

$$\text{cov}(E_{\text{av}}, E'_{\text{av}}) = \text{var}(E_{\text{av}})[1 - b\tau] + \cdots, \tag{23}$$

where b is a constant (depending only on ψ_T and V, but independent of M) given in (C5). Therefore the first-order serial correlation coefficient between E_{av} and E'_{av} equals

$${}^1\rho_{\text{av}} = 1 - b\tau + \cdots \tag{24}$$

and the kth order coefficient (using the Markovian assumption) is

$${}^k\rho_{\text{av}} = \exp(-bk\tau) \tag{25}$$

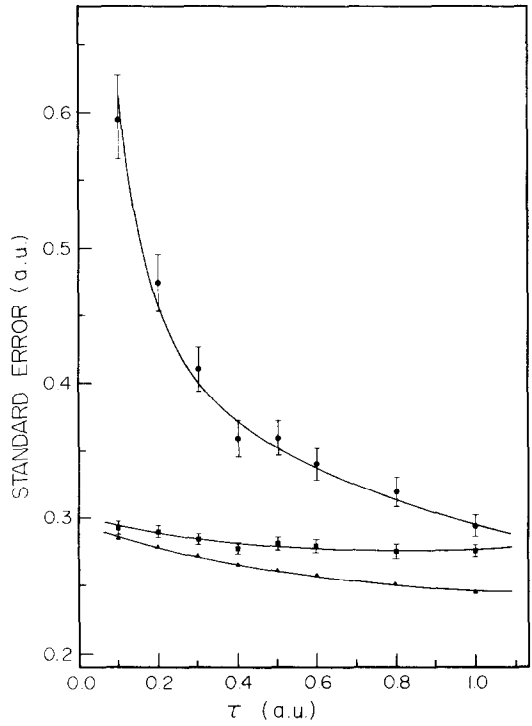

FIG. 1. Scaled standard error of various energy estimates (a.u.) versus τ (a.u.) for H_2 ground state: block mean, $(MK)^{1/2}$ s.e. (E_{bk}), Eq. (33) (●); iteration average of the local energy, $M^{1/2}$ s.e. (E_{av})) (■); local energy, s.e. (E_L) (▲).

or, more generally,

$$^k\rho_{av} = \sum B_i \exp(-b_i k\tau), \tag{26}$$

where expressions analogous to (19) and (20) hold for b_i and B_i.

Using (25) the variance of the grand-mean estimate E_{gm} equals

$$\text{var}^M(E_{gm}) = \frac{\text{var}(E_{av})}{K} \cdot \frac{1 + {}^1\rho_{av}}{1 - {}^1\rho_{av}} \tag{27}$$

or, to the smallest order of τ accuracy

$$\text{var}^M(E_{gm}) = [\text{var}(E_{av})/K] \cdot [2/b\tau] + \cdots \tag{28}$$

which follows immediately from (24), where the M superscript denotes the Markovian assumption.

The more general approach of (26) leads to

$$\text{var}(E_{gm}) = \text{var}(E_{av}) \left[1 + 2 \sum_{k=1} {}^k\rho_{av} \right] \Big/ K \tag{29}$$

which, referring to (18), can be expanded as

$$(\mathrm{var}(E_{\mathrm{av}})/K)\left[2/\tau \cdot \sum_i B_i/b_i\right] + \cdots. \tag{30}$$

The last expression may slightly differ from (28) in terms of the actual coefficient, but it is identically proportional to τ^{-1}. In the next section, we describe a procedure for constructing error bars for E_{gm}, which applies regardless of whether (27) is a sufficiently accurate approximation to (29) or not.

In summary, the statistical error of the grand-mean estimate depends on the variance of the individual iteration averages (22) and on the serial correlation of consecutive E_{av} values. The former depends predominantly on the quality of the guiding function and on C (B9), a component which arises from branching. An empirical study of the τ dependence of these effects, and of the overall grand-mean variance, is the subject of the following section.

Empirical Study

In this section we illustrate the main points of our paper by an empirical example. We will also describe how to reliably estimate the standard error of E_{gm}. Direct use of (29) is not recommended because the normally strong serial correlation of the iteration averages gives rise to a long sequence of nonzero serial correlation coefficients; due to statistical complications it would be very difficult to reliably estimate their infinite sum. Instead, we present a procedure which divides the sequence of E_{av}'s into several blocks of the same size, to produce a relatively accurate and simple estimate of the error bar of E_{gm}. We will contrast the results obtained from this procedure with those obtained without using blocks and assuming that the underlying sequence of E_{av}'s correspond first to the Markov, and then to the Yule models.

One first performs a run consisting of sufficiently many (several thousand) iterations, and the generated set of E_{av}'s is stored. Next one estimates the serial correlation coefficients ${}^{k}\rho_{\mathrm{av}}$. (We used ${}^{1}r_{\mathrm{av}}$ to ${}^{200}r_{\mathrm{av}}$, where ${}^{k}r$ is the usual estimate of ${}^{k}\rho$; see [10]. Note that all estimates may have a sizable statistical error and be negative from a certain order on, due to statistical bias; also, nonrandom-looking oscillations may persist throughout the sequence.) If the k_0th serial correlation coefficient is deemed to be practically negligible, divide the sample into N_{bk} blocks, each consisting of k_0 values of E_{av}. Then calculate the block averages $E_{\mathrm{bk},i}$ and estimate the variance and the first serial correlation coefficient of this sequence (all the higher serial correlations are equal to zero). The mean of the block averages $\underline{E}_{\mathrm{bk}}$ (identical to the grand-mean E_{gm}) is the ultimate estimate of $E(\tau)$, where

$$\underline{E}_{\mathrm{bk}} = \sum_{i=1}^{N_{\mathrm{bk}}} E_{\mathrm{bk},i}/N_{\mathrm{bk}}. \tag{31}$$

As the following model pertains to the series of block averages:

$$ {}^{1}\rho_{bk} = \rho_{bk}, \qquad {}^{k}\rho_{bk} = 0, \qquad k \neq 0, 1 \tag{32} $$

the standard error (s.e.) of the grand-mean can be easily estimated by

$$ \text{s.e.}(\widehat{E}_{bk}) = \{\text{var}(\widehat{E}_{bk})[1 + 2\rho_{bk}]/N_{bk}\}^{1/2}, \tag{33} $$

where the variance estimator is given by

$$ \text{var}(\widehat{E}_{bk}) = (N_{bk} - 1)^{-1}\left[\sum_{i=1}^{N_{bk}} (E_{bk,i}^2) - N_{bk}^{-1}\left(\sum_{i=1}^{N_{bk}} E_{bk,i}\right)^2\right] \tag{34} $$

Note that the serial correlation of the blocks, if ignored, leads to an underestimation of the standard error. One estimates ρ_{bk} from the variance and covariance of the sample of blocks; the relative error of this estimate of (33) is approximately $\sqrt{(6/N_{bk})}$.

This procedure has the advantage of yielding a large number of simply correlated blocks of data, and thus it is more efficient and less prone to bias than the technique of creating significantly fewer huge blocks of data, N'_{bk}, which may be treated as statistically independent observations (relative error approximately $\sqrt{(2/N'_{bk})}$).

It is instructive to estimate the standard error without using blocks by assuming that the sequence of E_{av}'s conforms to a simple (autoregressive) time series model. We tried the first two simplest possibilities: Markov, with one parameter, and Yule [10], with two. Equation (27) holds for the former; for the latter,

$$ \text{var}^{Y}(E_{gm}) = \text{var}(E_{av})(1 + {}^{1}\rho_{av})(1 - 2({}^{1}\rho_{av})^2 + {}^{2}\rho_{av})/[(1 - {}^{1}\rho_{av})(1 - {}^{2}\rho_{av})]\, K, \tag{35} $$

where Y denotes the Yule assumption. For a related application of these models in statistical mechanics, see [11].

The Markov model (in spite of not being totally adequate otherwise) may be used to facilitate the determination of k_0. When the serial correlation coefficient ${}^{k}\rho_{av}$ is practically zero (that is, k is sufficiently large) the standard error of our estimate is given by [10]:

$$ \text{s.e.}({}^{k}r_{av}) = \{[(1 + ({}^{1}\rho_{av})^2)/(1 - ({}^{1}\rho_{av})^2)]/M\}^{1/2}. \tag{36} $$

This equation provides a rule-of-thumb estimate of the size of the random oscillations in the sequence of serial correlation coefficients. A reasonable stopping rule based on this is as follows: let k_0 be the point at which the sequence of ${}^{k}r_{av}$ values is consistently less than 2.5 times s.e.$({}^{k}r_{av})$, as given by (36). Another generally more conservative but simpler rule is to choose the point at which the sequence first becomes negative. (To reach a reasonably small statistical error of the

variance estimate, one should always try *not* to be too conservative and to use as small a value of k_0 as possible; the potential bias of the final estimate is insignificant when compared to the statistical error itself. (Any prescription for finding k_0 should not be applied blindly; we recommend verifying the reasonableness of the final choice of k_0 by inspection of the ${}^k r$ sequence itself.)

As an empirical example, we generated a sample which consisted of $K = 10{,}000$ iterations on $M = 1000$ configurations for a wide range of τ values using the algorithm of [9] applied to H_2. For each iteration the average (E_{av}) and standard error of the E_L's was tabulated. The average standard error over the K iterations was then determined (s.e. (E_L)). Also, the average of the sample of the E_{av}'s (E_{gm}) and its standard error, ignoring the serial correlation, (s.e. (E_{av})), were computed. Next, the sample was divided in N_{bk} blocks, each of size k_0. Finally, the standard error of the mean of the block averages ($\underline{E}_{bk} = E_{gm}$) was computed (s.e. ($\underline{E}_{bk}$)), using (33). The standard error results appear in Fig. 1. To faciliate a meaningful

TABLE I

Grand Mean Energy and Standard Error for H_2 Ground State versus Time Step[a]

τ	$\underline{E}_{bk}$[b]	k_0[c]	s.e. ($\underline{E}_{bk}$)[d] [$\times 10^3$]	s.e. ${}^Y(E_{gm})$[e] [$\times 10^3$]	s.e. ${}^M(E_{gm})$[f] [$\times 10^3$]
0.1	−1.174(2)	14	1.90(9)	1.83(6)	1.59(4)
0.2	−1.172(2)	9	1.48(6)	1.45(3)	1.30(3)
0.3	−1.168(1)	8	1.30(6)	1.26(3)	1.15(2)
9.4	−1.164(1)	7	1.14(3)	1.14(3)	1.05(2)
0.5	−1.159(1)	6	1.14(3)	1.11(3)	1.03(2)
0.6	−1.154(1)	6	1.08(3)	1.08(3)	0.99(2)
0.8	−1.145(1)	6	1.01(3)	1.01(3)	0.94(2)
1.0	−1.137(1)	4	0.92(3)	0.92(3)	0.90(2)

Model[g]

$$\underline{E}_{bk} = E_0 + a\tau^2 + b\tau^4$$

$$E_0 = -1.1735(9), \qquad a = 0.063(5), \qquad b = -0.266(5)$$

$$\mathrm{var}(\underline{E}_{bk}) = a/\tau + b + c\tau$$

$$a = 2.46(8) \times 10^{-7}, \qquad b = 9.29(3) \times 10^{-7}, \qquad c = -3.29(2) \times 10^{-7}$$

[a] All quantities are in atomic units. The estimates are uncertain in the last decimal place quoted; the standard error appears in parentheses.

[b] Each simulation uses the algorithm of [9] for ($K =$) 10,000 iterations on an iniial list of ($M =$) 1000 configurations. The averages and standard errors are for blocks of iteration averages, Eq. (7), each block of size k_0, using Eqs. (31)–(34). $E_0^{\text{exact}} = -1.1745$ a.u. and ψ_T is ψ_I taken from [1], with the nuclei separation of 1.401 a.u.

[c] Block size.

[d] Standard error derived from blocking the data; Eq. (33).

[e] Standard error derived from the Yule model; square root of Eq. (35).

[f] Standard error derived from the Markov model; square root of Eq. (27).

[g] Weighted polynomial regression.

comparison of these quantities, we plotted $[(MK)^{1/2}\,\text{s.e.}\,(\underline{E}_{bk})]$ and $[M^{1/2}\,\text{s.e.}\,(E_{av})]$, quantities on the same scale as $\underline{\text{s.e.}}\,(E_L)$. The values of E_{gm} and k_0 for each value of τ appear in Table I.

In Fig. 1 the fact that $[M^{1/2}\,\text{s.e.}\,(E_{av})]$ is consistently larger than $\underline{\text{s.e.}}\,(E_L)$ is due to the branching (see discussion of (22)). As expected, this effect clearly becomes important as τ increases. The consistently larger values for $[(MK^{1/2}\,\text{s.e.}\,(\underline{E}_{bk})]$ than for $[M^{1/2}\,\text{s.e.}\,(E_{av})]$ are due to the serial correlation of the E_{av}'s, which was ignored in the latter quantity. Finally, the behavior of $[(MK)^{1/2}\,\text{s.e.}\,(\underline{E}_{bk})]$ at small τ is $\tau^{-1/2}$; at large τ one sees the effect of higher τ order terms in the expression for the serial correlation. All this is in accord with (30). The results of a regression fit versus τ appear in Table I.

As an illustration of the use of (36) to estimate k_0, we display, in Fig. 2, our estimates of the first few serial correlation coefficients (τ equals 0.2). The cutoff value of nine, computed from the formula, yields more blocks (hence better precision for the ultimate estimate of the variance) than using the more conservative cutoff value of seventeen (the first occurrence of a negative correlation coefficient). The seemingly nonrandom oscillations observed beyond $k=9$ are consistently (allowing for an odd outlier or two) within 2.5 times the standard error of

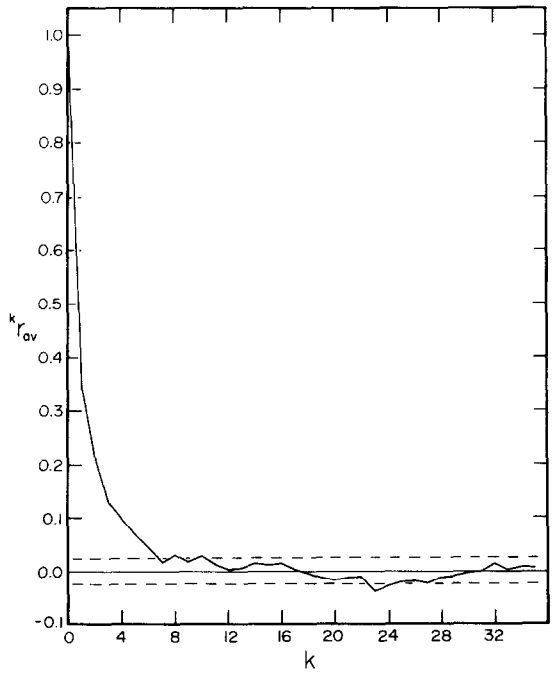

FIG. 2. Correlogram obtained from a sample of 10,000 iteration averages for H_2 ground state generated at $\tau=0.2$ a.u. The dotted lines denote 2.5 times the standard error of the serial correlation coefficients, Eq. (36), at large k.

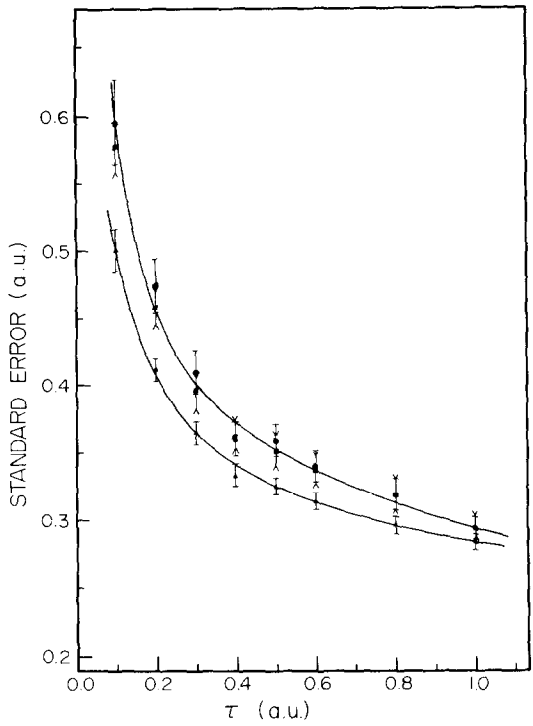

FIG. 3. Scaled standard error of grand mean energy (a.u.) versus τ (a.u.) for H_2 ground state: using blocks, $(MK)^{1/2}$ s.e. $(\underline{E}_{bk})$, Eq. (33) (●); using Yule autoregressive model, $(MK)^{1/2}$ s.e. ${}^{Y}(E_{gm})$, Eq. (35) (■); using Markov autoregressive model, $(MK)^{1/2}$ s.e. ${}^{M}(E_{gm})$, Eq. 27 (▲).

${}^{k}r_{av}$ given by (36). Such oscillations are both typical and expected; even a completely random set of observations will have a correlogram with similar behavior. (For example, see Fig. 3.1 in [12].)

In Fig. 3 we contrast s.e. $(\underline{E}_{bk})$ with the standard error of E_{gm} assuming that the series of E_{av}'s obey the Markov (27) and the Yule (35) models. The Yule value provides a consistently good estimate of the standard error for the entire range of τ values, but even this model will break down for τ smaller than 0.1. The Markov model consistently underestimates the standard error for our range of τ values. Thus we cannot recommend using the Markovian assumption in the estimation of the serial correlation (such as for the optimization of DMC [7]), although the model itself is sufficient to predict the τ^{-1} behavior of the leading term in the statistical variance (28).

APPENDIX A: COVARIANCE OF CONSECUTIVE LOCAL ENERGIES

From Eqs. (12) and (13)

$$\begin{aligned}
&\operatorname{cov}(E_L(\mathbf{R}_i), E_L(\mathbf{R}_i')) \\
&\quad = \iint_{\mathbf{R}_-} (E_L(\mathbf{R}_i) - E_0)(E_L(\mathbf{R}_i') - E_0)\, g(\mathbf{R}_i, \mathbf{R}_i')\, d\mathbf{R}_i\, d\mathbf{R}_i' \\
&\qquad + \iint_{\mathbf{R}_+} (E_L(\mathbf{R}_i) - E_0)(E_L(\mathbf{R}_i') - E_0)(1 - \tau(E_L(\mathbf{R}_i) - E_0))\, g(\mathbf{R}_i, \mathbf{R}_i')\, d\mathbf{R}_i\, d\mathbf{R}_i' \\
&\qquad + \iint_{\mathbf{R}_+} (E_L(\mathbf{R}_i) - E_0)(E_L(\mathbf{R}_i') - E_0)\, \tau(E_L(\mathbf{R}_i) - E_0)\, g^-(\mathbf{R}_i, \mathbf{R}_i')\, d\mathbf{R}_i\, d\mathbf{R}_i' \\
&\qquad + \cdots, \qquad \text{(A1)}
\end{aligned}$$

where $\mathbf{R}_+$ is the region of the "excessive" $\mathbf{R}$ values (the complement of $\mathbf{R}_-$), and the $\mathbf{R}'$ integrals are over all space. The third term (with pdf $g^-(\mathbf{R}_i, \mathbf{R}_i')$) represents the covariance between the local energy evaluated at $\mathbf{R}_i$, a configuration in the present iteration which has been deleted for the next iteration, and the local energy evaluated at $\mathbf{R}_i'$, a configuration which has been duplicated.

Substitute $\mathbf{R}_i^a$ of (5) for $\mathbf{R}_i'$ in the first two terms of (A1). Now, the joint distribution $g(\mathbf{R}_i, \mathbf{R}_i')\, d\mathbf{R}_i\, d\mathbf{R}_i'$ becomes $f(\mathbf{R}_i)\, h(\chi_i)\, d\mathbf{R}_i\, d\chi_i$, where h is the (usually normal) standardized distribution of independent components. Furthermore, from (5) it follows that

$$\begin{aligned}
\int E_L(\mathbf{R}_i^a)\, h(\chi_i)\, d\chi_i &= \int \{E_L(\mathbf{R}_i) + (\tau \mathbf{F}(\mathbf{R}_i) + \tau^{1/2}\chi_i) \cdot \nabla E_L(\mathbf{R}_i) \\
&\qquad + \tfrac{1}{2}[(\tau \mathbf{F}(\mathbf{R}) + \tau^{1/2}\chi_i) \cdot \nabla]^2 E_L(\mathbf{R}_i) + \cdots\}\, h(\chi_i)\, d\chi_i \\
&= E_L(\mathbf{R}_i) + \tau \mathbf{F}(\mathbf{R}_i) \cdot \nabla E_L(\mathbf{R}_i) + \tfrac{1}{2}\tau \nabla^2 E_L(\mathbf{R}_i) + \cdots \qquad \text{(A2)}
\end{aligned}$$

Substituting (A2) into (A1) yields

$$\begin{aligned}
&\operatorname{cov}(E_L(\mathbf{R}_i), E_L(\mathbf{R}_i')) \\
&\quad = \int [(E_L(\mathbf{R}) - E_0)^2 + (E_L(\mathbf{R}) - E_0)\, \tau \mathbf{F}(\mathbf{R}) \cdot \nabla E_L(\mathbf{R}) \\
&\qquad + \tfrac{1}{2}\tau (E_L(\mathbf{R}) - E_0)\, \nabla^2 E_L(\mathbf{R})]\, f(\mathbf{R})\, d\mathbf{R} + \tau \int_{\mathbf{R}_+} (E_L(\mathbf{R}) - E_0)^3 f(\mathbf{R})\, d\mathbf{R} \\
&\qquad + \tau \int_{\mathbf{R}_+} (E_L(\mathbf{R}) - E_0)^2 (E_L(\mathbf{R}_i') - E_0)\, g^-(\mathbf{R}_i, \mathbf{R}_i')\, d\mathbf{R}\, d\mathbf{R}' + \cdots. \qquad \text{(A3)}
\end{aligned}$$

From Green formulas and from (2)

$$\int [(E_L(\mathbf{R})-E_0)^2+(E_L(\mathbf{R})-E_0)\,\tau\mathbf{F}(\mathbf{R})\cdot\nabla E_L(\mathbf{R})+\tfrac{1}{2}\tau(E_L(\mathbf{R})-E_0)\,\nabla^2 E_L(\mathbf{R})$$

$$-\tfrac{1}{2}\tau(E_L(\mathbf{R})-E_0)^3]\,f(\mathbf{R})\,d\mathbf{R}$$

$$=\int (E_L(\mathbf{R})-E_0)^2(1-\tfrac{1}{2}\tau\nabla\cdot\mathbf{F}(\mathbf{R})+\tfrac{1}{4}\tau\nabla^2-\tfrac{1}{2}\tau(E_L(\mathbf{R})-E_0))f(\mathbf{R})\,d\mathbf{R}$$

$$-\tfrac{1}{2}\tau\int[\nabla E_L(\mathbf{R})]^2 f(\mathbf{R})\,d\mathbf{R}$$

$$=\int (E_L(\mathbf{R})-E_0)^2 f(\mathbf{R})\,d\mathbf{R}-\tfrac{1}{2}\tau\int[\nabla E_L(\mathbf{R})]^2 f(\mathbf{R})\,d\mathbf{R}. \tag{A4}$$

The operator range of ∇ is restricted to within the square brackets in (A3) and (A4).

Upon substitution of (A4) into (A3)

$$\operatorname{cov}(E_L(\mathbf{R}_i),E_L(\mathbf{R}_i'))$$

$$=\int (E_L(\mathbf{R})-E_0)^2 f(\mathbf{R})\,d\mathbf{R}-\tfrac{1}{2}\tau\int|E_L(\mathbf{R})-E_0|^3 f(\mathbf{R})\,d\mathbf{R}$$

$$-\tau\int_{\mathbf{R}_+}(E_L(\mathbf{R})-E_0)^2(E_0-E_L(\mathbf{R}'))\,g^-(\mathbf{R},\mathbf{R}')\,d\mathbf{R}\,d\mathbf{R}'$$

$$-\tfrac{1}{2}\tau\int[\nabla E_L(\mathbf{R})]^2 f(\mathbf{R})\,d\mathbf{R}$$

$$+\cdots. \tag{A5}$$

The first factor is the variance of $E_L(\mathbf{R})$. Now (14) follows immediately from this upon defining the following constant

$$a\equiv\operatorname{var}(E_L(\mathbf{R}))^{-1}\left\{\frac{1}{2}\int|E_L(\mathbf{R})-E_0|^3 f(\mathbf{R})\,d\mathbf{R}+\frac{1}{2}\int[\nabla E_L(\mathbf{R})]^2 f(\mathbf{R})\,d\mathbf{R}\right.$$

$$\left.+\int_{\mathbf{R}_+}(E_L(\mathbf{R})-E_0)^2(E_0-E_L(\mathbf{R}'))\,g^-(\mathbf{R},\mathbf{R}')\,d\mathbf{R}\,d\mathbf{R}'\right\}. \tag{A6}$$

APPENDIX B: DERIVATION OF $\operatorname{var}(E_{\mathrm{av}})$

From our description of steps (c) and (d) above (12), to the first order accuracy in τ, the expected number of duplicate pairs created in one iteration is equal to

$$M\tau\int_{\mathbf{R}_-}(E_0-E_L(\mathbf{R}))f(\mathbf{R})\,d\mathbf{R}. \tag{B1}$$

A pair of originally identical values of $E_L(\mathbf{R}_i)$ has one iteration later a covariance equal to

$$\operatorname{cov}^-(E_L(\mathbf{R}'_{i1}), E_L(\mathbf{R}'_{i2}))$$

$$= \iint (E_L(\mathbf{R}'_{i1}) - E')(E_L(\mathbf{R}'_{i2}) - E')\, g'(\mathbf{R}'_{i1}, \mathbf{R}'_{i2})\, d\mathbf{R}'_{i1}\, d\mathbf{R}'_{i2} \tag{B2}$$

where $\mathbf{R}'_{i1}$ and $\mathbf{R}'_{i2}$ represent the two values of a duplicate configuration gone through one complete iteration (steps (b)–(d)). In (B2) g' is the joint distribution function of $\mathbf{R}'_{i1}$ and $\mathbf{R}'_{i2}$, E' is the expected value of $E_L(\mathbf{R}'_{i1})$ (equal to that of $E_L(\mathbf{R}'_{i2})$), and the prime superscript implies averaging with respect to g'. These configurations were created in step (c) by duplicating $\mathbf{R}_i$, a member of the set of "deficient configurations." Furthermore, for each $\mathbf{R}'_{i1}$ and $\mathbf{R}'_{i2}$ we use independent χ_{i1} and χ_{i2}:

$$\mathbf{R}'_{i1} = \mathbf{R}_i + \tau \mathbf{F}(\mathbf{R}_i) + \tau^{1/2} \chi_{i1} \tag{B3}$$

and the analog for $\mathbf{R}'_{i2}$. Thus $g'(\mathbf{R}'_{i1}, \mathbf{R}'_{i2})\, d\mathbf{R}'_{i1}\, d\mathbf{R}'_{i2}$ is equivalent to $f^-(\mathbf{R}_i)\, h(\chi_{i1})\, h(\chi_{i2})\, d\mathbf{R}_i\, d\chi_{i1}\, d\chi_{i2}$, where f^- is the distribution of deficient configurations (12) from which the parent configuration $\mathbf{R}_i$ was drawn. (B2) is now equal to

$$\operatorname{cov}^-(E_L(\mathbf{R}'_{i1}), E_L(\mathbf{R}'_{i2}))$$

$$= \iiint (E_L(\mathbf{R}'_{i1}) - E'_L)(E_L(\mathbf{R}'_{i2}) - E'_L) f^-(\mathbf{R}_i)\, h(\chi_{i1})\, h(\chi_{i2})\, d\mathbf{R}_i\, d\chi_{i1}\, d\chi_{i2}. \tag{B4}$$

We may now integrate over χ_{i1} and χ_{i2} as in (A2).

When calculating the corresponding variance of $E_L(\mathbf{R}'_{i1})$ and $E_L(\mathbf{R}'_{i2})$, we obtain

$$\operatorname{var}^-(E_L(\mathbf{R}'_i)) = \iint (E_L(\mathbf{R}'_i) - E'_L)^2 f^-(\mathbf{R}_i)\, h(\chi_i)\, d\mathbf{R}_i\, d\chi_i. \tag{B5}$$

The expressions for cov^- and var^- are thus identical except for the integral of the cross term involving $\chi_{i1}\chi_{i2}$ and χ_i^2, respectively, which integrates to zero in the covariance case but contributes to the variance. Thus we get

$$\operatorname{cov}^-(E_L(\mathbf{R}'_{i1}), E_L(\mathbf{R}'_{i2}))$$

$$\operatorname{var}^-(E_L(\mathbf{R}'_i)) - \tau \int (\nabla E_L(\mathbf{R}_i))^2 f^-(\mathbf{R}_i)\, d\mathbf{R}_i + \cdots. \tag{B6}$$

Finally [by the argument leading to (16) and (17)] the correlation for a duplicate configuration, k iterations later, is given by (suppressing the index i):

$${}^k\rho^- = \left\{ \exp\left[-\tau \int (\nabla E_L(\mathbf{R}))^2 f^-(\mathbf{R})\, d\mathbf{R} / \operatorname{var}^-(E_L(\mathbf{R}')) \right] \right\}^k \equiv \{\exp[-c\tau]\}^k. \tag{B7}$$

The cumulative amount of covariance contributed by a *single* duplicate pair to the overall sum (M configurations advanced through K iterations) equals

$$\operatorname{var}^-(E_L(\mathbf{R}')) \cdot [1 + {}^1\rho^- + {}^2\rho^- + {}^3\rho^- + \cdots] \approx \operatorname{var}^-(E_L(\mathbf{R}'))/(c\tau), \tag{B8}$$

where ${}^1\rho^-$ is the first-order serial correlation coefficient from (B7). Equations (B7) and the right-hand side of (B8) rest on the Markovian assumption (the τ^{-1} behavior is *general*). The total amount of covariance introduced by *all* duplicates throughout K iterations is thus equal to $K\cdot(\mathrm{B1})\cdot(\mathrm{B8})$. Therefore in a single-iteration (divide by K) this contributes to the variance of $\sum_{i=1} E_L(\mathbf{R}_i)$ an additional factor of $2\cdot(\mathrm{B1})\cdot(\mathrm{B8})$. Finally, dividing by M^2 gives the contribution to the variance of E_{av}. This results in the extra $2C/M$ of (22), where C is equal to

$$C \equiv \int_{\mathbf{R}_-} (E_0 - E_L(\mathbf{R}))f(\mathbf{R})\,d\mathbf{R}\cdot \mathrm{var}^-(E_L(\mathbf{R}')/c + \cdots. \tag{B9}$$

APPENDIX C: COVARIANCE OF E_{av}

The variance of $\sum_{i=1}^{M} E_L(\mathbf{R}_i)$, namely,

$$\mathscr{E}\left[\sum_{i=1}^{M}(E_L(\mathbf{R}_i)-E_0)\cdot\sum_{j=1}^{M}(E_L(\mathbf{R}_j)-E_0)\right], \tag{C1}$$

where $\mathscr{E}$ denotes the expected value with respect to the *joint* distribution of all M configurations, must remain the same throughout the iterations, once the stationary stage of the simulation has been reached. This means that (C1) is equal to

$$\mathscr{E}\left[\sum_{i=1}^{M}(E_L(\mathbf{R}_i')-E_0)\cdot\sum_{j=1}^{M}(E_L(\mathbf{R}_j')-E_0)\right]. \tag{C2}$$

From the equality of (C1) and (C2) and after using (A2) [note that the terms of the type $\tau^{1/2}\chi_i\cdot\nabla E_L(\mathbf{R}_i)$ do contribute when $i=j$], one can prove that

$$2\mathscr{E}\left[\sum_{i=1}^{M}\left(\mathbf{F}(\mathbf{R}_i)\cdot\nabla E_L(\mathbf{R}_i)+\frac{1}{2}\nabla^2 E_L(\mathbf{R}_i)\right)\cdot\sum_{j=1}^{M}(E_L(\mathbf{R}_j)-E_0)\right]$$
$$=-\mathscr{E}\left[\sum_{i=1}^{M}(\nabla E_L(\mathbf{R}_i)^2\right]. \tag{C3}$$

From this equation it further follows that

$$\mathscr{E}\left[\sum_{i=1}^{M}(E_L(\mathbf{R}_i')-E_0)\cdot\sum_{j=1}^{M}(E_L(\mathbf{R}_j)-E_0)\right]$$
$$=\mathscr{E}\left[\sum_{i=1}^{M}(E_L(\mathbf{R}_i)-E_0)\cdot\sum_{j=1}^{M}(E_L(\mathbf{R}_j)-E_0)\right]+\mathscr{E}\left[\tau\sum_{i=1}^{M}\left(\mathbf{F}(\mathbf{R}_i)\cdot\nabla E_L(\mathbf{R}_i)\right.\right.$$
$$\left.\left.+\frac{1}{2}\nabla^2 E_L(\mathbf{R}_i)\right)\cdot\sum_{j=1}^{M}(E_L(\mathbf{R}_j)-E_0)\right]+\cdots$$
$$=\mathrm{var}\left(\sum_{i=1}^{M}E_L(\mathbf{R}_i)\right)-\frac{1}{2}\tau\mathscr{E}\left[\sum_{i=1}^{M}(\nabla E_L(\mathbf{R}_i))^2\right]+\cdots. \tag{C4}$$

Now (23) follows immediately upon defining the constant:

$$b \equiv \frac{1}{2} \int (\nabla E_L(\mathbf{R}_i))^2 f(\mathbf{R})\, d\mathbf{R} / [M \cdot \mathrm{var}(E_{\mathrm{av}})]. \tag{C5}$$

REFERENCES

1. P. J. REYNOLDS, D. M. CEPERLEY, B. J. ALDER, AND W. A. LESTER, JR., *J. Chem. Phys.* **77**, 5593 (1982).
2. R. J. HARRISON AND N. C. HANDY, *Chem. Phys. Lettr.* **113**, 257 (1985).
3. P. J. REYNOLDS, M. DUPUIS, AND W. A. LESTER, JR., *J. Chem. Phys.* **82**, 1983 (1985).
4. R. N. BARNETT, P. J. REYNOLDS, AND W. A. LESTER, JR., *J. Chem. Phys.* **82**, 2700 (1985).
5. N. C. HANDY, R. J. HARRISON, P. J. KNOWLES, AND H. F. SCHAEFER III, *J. Phys. Chem.* **88**, 4852 (1984).
6. H. MÜLLER-KRUMBHAAR AND K. BINDER, *J. Statist. Phys.* **8**, 1 (1973).
7. J. VRBIK AND S. M. ROTHSTEIN, *Int. J. Quant. Chem.* **29**, 416 (1986).
8. J. VRBIK, *J. Phys. A* **18**, 1327 (1985).
9. J. VRBIK AND S. M. ROTHSTEIN, *J. Comput. Phys.* **63**, 130 (1986).
10. M. G. KENDALL AND A. STUART, *The Advanced Theory of Statistics, Vol. 3* (Hafner, New York, 1968), Chap. 48.
11. E. B. SMITH AND B. H. WELLS, *Mol. Phys.* **53**, 701 (1984).
12. M. G. KENDALL, *Time-Series* (Hafner, New York, 1976).