JOURNAL OF COMPUTATIONAL PHYSICS **89**, 432–438 (1990)

Statistical Error Methods in Computer Simulations

JUAN J. MORALES AND MARÍA J. NUEVO

Departamento de Fisica, Facultad de Ciencias,
Universidad de Extremadura,
06071 Badajoz, Spain

AND

LUIS F. RULL

Departamento de Física Atómica, Molecular y Nuclear,
Facultad de Fisica, Universidad de Sevilla,
Apartado 1065, 41080 Seville, Spain

Received January 30, 1989

Two methods of studying the statistical error in the sequences of data obtained by computer simulation are compared. The first method divides the sequence into blocks whose length is selected graphically by means of the "statistical inefficiency." The second method uses the autocorrelation function of all the values obtained and analytically calculates its convergence by means of the "correlation length." The general relationship between the two parameters is found mathematically and is in good agreement with the experimental data obtained by molecular dynamics simulation in the melting zone when a very accurate algorithm is used. As a consequence, the analytical method is more accurate than the graphical method. © 1990 Academic Press, Inc.

I. INTRODUCTION

One of the more serious problems of computer simulation is the estimation of the statistical errors when the data series is strongly correlated. To overcome this problem the common procedure is to divide the series into a number of subseries of a length such that one expects the subaverages to be randomly distributed, and then the error in the mean of the complete series can be estimated from the standard deviation of the mean of those subaverages. The length of each subseries must be larger than the correlation time to ensure that averages from these subseries are uncorrelated. Thus the basic problem in determining the statistical errors of a finite correlated series is to choose the number of intervals into which the series should be divided.

There are different methods of choosing the appropriate number of subseries. One is the earlier method due to Friedberg and Cameron [1] that they used to test the Monte Carlo (MC) simulation of a small Ising lattice. The analysis of the error

0021-9991/90 $3.00

is basically the following: If successive configurations are correlated with respect to a variable x, then the variance, $\sigma^2(x_i)$, $i = 1 \cdots n \cdots N$, will rise with increasing n until it approaches a limiting value. The leveling off of the variance signifies that n has become so large that there is no correlation from block to block, so that on further agglutination of blocks the variances are Gaussian. The "statistical inefficiency" (SI) is defined as the limiting ratio of the observed variance of the long-term averages to their expected (Gaussian) variance. Thus the statistical inefficiency is simply related to the correlation time. This method has been used in a recent application by Fichman *et al.* [2], using molecular dynamics (MD) simulation of molecular liquid mixtures.

Another kind of method has been proposed by Smith and Wells [3]. They described a statistical procedure wherein the nature of the correlation between terms in the data sequence can be evaluated using the autocorrelation function. Treating the data sequence as an autoregressive process, they found the series to be first-order because, in general, an assumption of second-order behavior led to estimates of the error that were significantly too low.

A much simpler generalization of this method was used by Straatsma *et al.* [4]. With their procedure it is not necessary to assume any particular type of correlated behavior in the data sequence, thus avoiding the difficult analysis of the partial autocorrelation and at the same time utilizing all the information contained in the autocorrelation function of the correlated series calculated by means of the "correlation length," τ. With their procedure of subaverages over intervals, they found that when the interval size is too small the averages cannot be considered uncorrelated, whereas for large intervals, the small number of intervals gives a statistically inaccurate error.

In Section II we describe the theoretical background to calculating the error in the mean, using the Friedberg and Cameron and the Straatsma *et al.* methods, obtaining the theoretical relation between SI and τ. In Section III, we obtain SI graphically and τ analytically from the values obtained by MD simulation in the melting zone, using two different algorithms. Finally, the conclusions about the concordance between the theoretical result (Sect. II) and experimental results (Sect. III) are shown in Section IV.

II. Statistical Method

The results of an MD computer simulation are in the form of N values of a certain property x_i. Each value is separated from the next by an interval which is a multiple of the time step h. We have to determine whether or not there exists a correlation between the values obtained.

(a) The Friedberg and Cameron method consists in dividing the N numbers into n_b blocks, each with n successive values

$$N = n \times n_b. \tag{1}$$

The value of the mean (subaverage) for each block is

$$\bar{x}_b = (1/n) \sum_{i=1}^{n} x_i \tag{2}$$

and the total mean of the subaverages is

$$\bar{x} = (1/n_b) \sum_{b=1}^{n_b} \bar{x}_b \tag{3}$$

Equation (3) is equivalent to the mean of the whole set of values

$$\bar{x} = (1/N) \sum_{i=1}^{N} x_i. \tag{4}$$

Using the central limit theorem, the variance in the means, if the subaverages from blocks of length n give independent estimates of the mean, is

$$\sigma^2(x) = \frac{\sigma^2(\bar{x}_b)}{n_b}, \tag{5}$$

where the variance of the means is

$$\sigma^2(\bar{x}_b) = (1/n_b) \sum_{b=1}^{n_b} (x_b - x)^2. \tag{6}$$

If all the values were independent then $\bar{x}_b \equiv x_i$ and $n_b = N$ and Eqs. (5) and (6) reduce respectively to

$$\sigma^2(x) = \frac{\sigma^2(x_i)}{N} \tag{5a}$$

and

$$\sigma^2(x_i) = (1/N) \sum_{i=1}^{N} (x_i - \bar{x})^2. \tag{6a}$$

The statistical inefficiency is defined as

$$\mathrm{SI} = \lim_{n \to \infty} \frac{n\sigma^2(\bar{x}_b)}{\sigma^2(x_i)} \tag{7}$$

in order to determine the appropriate block length n.

The limiting process is done by plotting $n\sigma^2(\bar{x}_b)/\sigma^2(x_i)$ against $n^{1/2}$ until the ordinates fluctuate about some given value which is taken as the value of SI. We should increase the number of steps by this factor of SI to compensate for the fact that successive states are correlated [2].

The error in the mean is given by

$$\sigma^2(\bar{x}) = \frac{\sigma^2(x_i)\,SI}{N} = \frac{\sigma^2(x_i)}{n}, \tag{8}$$

where

$$n = N/SI \tag{9}$$

is the effective number of independent data points.

(b) The Straatsma *et al.* method of studying the correlation between terms in the data series uses the autocorrelation function, at lag k, defined by

$$\hat{r}_k = \sum_{i=k+1}^{N} (x_i - \bar{x})(x_{i-k} - \bar{x}) \Big/ \sum_{i=1}^{N} (x_i - \bar{x})^2, \tag{10}$$

where $\bar{x}$ is calculated as in Eq. (4).

To estimate the error in the mean, they used the general expression

$$\sigma^2(\bar{x}) = \frac{\sigma^2(x_i)}{N}\left[1 + 2\sum_{k=1}^{N-1} \hat{r}_k\right], \tag{11}$$

where $\sigma^2(x_i)$ is defined by Eq. (6a), and $\tau = \sum_{k=1}^{N-1} \hat{r}_k$ is called the correlation length of the series and is obtained analytically by increasing the order in k, until $\hat{r}_k$ takes values around zero.

If all N values are uncorrelated the correlation length is zero and Eq. (11) reduces to Eq. (5a) as expected; but if the data points are correlated, the effective number of independent values will be

$$n = N/(1 + 2\tau) \tag{12}$$

and Eq. (11) is equivalent to Eq. (8).

Now the relation between the correlation length and the statistical inefficiency is obtained from Eqs. (9) and (12),

$$\mathrm{SI} = 1 + 2\tau. \tag{13}$$

III. Applications and Results

We have applied both methods to the microcanonical ensemble MD(EVN) for a bidimensional system of 576 particles at $kT/\varepsilon = 0.7$ and a density $\rho\sigma^2 = 0.84$ in the melting zone where the correlations have a long-range trend. This system was obtained by density scaling from the liquid system at $\rho\sigma^2 = 0.75$. The molecular pair potential was a Lennard–Jones truncated at $r_c = 2.5 \times 2^{1/6}\sigma \approx 2.8\sigma$. To integrate the

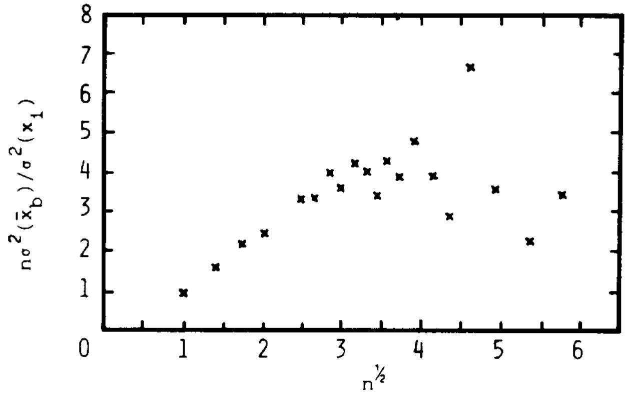

FIG. 1. The plateau gives the statistical inefficiency as in Refs. [1–2], using the Verlet algorithm.

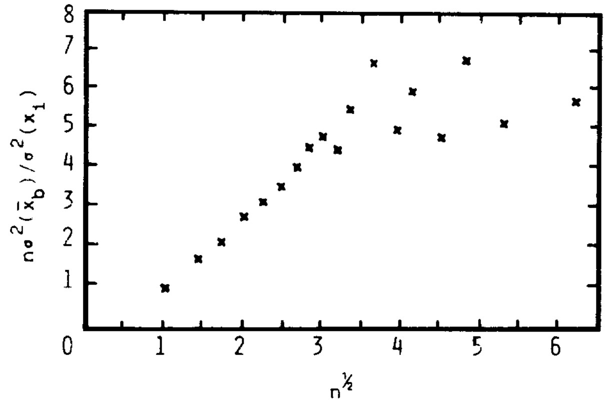

FIG. 2. The same as Fig. 1, but with the Toxvaerd algorithm.

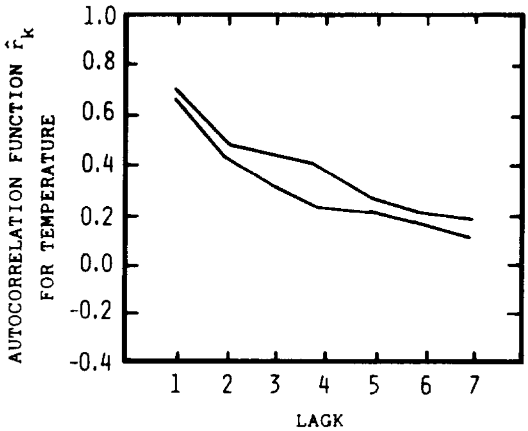

FIG. 3. Autocorrelation function, $\hat{r}_k$, for temperature at lag k, using the Verlet (lower) and Toxvaerd (upper line) algorithms.

equations of motion of the particles we used the Verlet [5] and the Toxvaerd [6] algorithms. The MD time step was $h = 0.005\ (m\sigma^2/\varepsilon)^{1/2}$.

The property we chose to test the statistical methods was the temperature. The averaging was done each 800 h during the free evolution of the system until 140,000 h and 115,000 h with the Verlet and Toxvaerd algorithms, respectively. More details about the simulation can be found in Ref. [7].

Once the data series have been obtained from our MD simulation, the correlation can be determined in two ways by using the two methods.

(a) With the Friedberg and Cameron method, SI can be found graphically. Plotting $n\sigma^2(\bar{x}_b)/\sigma^2(x_i)$ against $n^{1/2}$ for several values of n, we obtained the statistical inefficiency for temperature when the Verlet (Fig. 1) and Toxvaerd algorithms (Fig. 2) were used. These figures show the qualitative behavior of the correlation between the data series. Both begin with a linear rise of $n\sigma^2(\bar{x}_b)/\sigma^2(x_i)$ until $n^{1/2} \approx 2.5$ and $n^{1/2} \approx 3$, respectively, when fluctuation around a limiting value takes place. These values, which we have as the SI in the system, are about 4 in Fig. 1 and about 6 in Fig. 2.

(b) Using the Straatsma *et al.* method, we obtain the τ value analitycally as we saw in Section II(b). In Fig. 3, the autocorrelation function $\hat{r}_k$ for temperature is plotted versus lag k, up to seventh order in k, with the Toxvaerd algorithm (upper line) and the Verlet algorithm (lower line). The two lines fall to zero asymptotically: the fall of $\hat{r}_k$ with the first and second order in k are very similar in both algorithms, but for higher order in k the drop with the Verlet algorithm is greater.

Table I shows the results with the two algorithms; with N the total number of data obtained each $800h$ in our MD simulation. The values for the correlation length obtained analytically up to seventh order in k are shown under $1+2\tau$ as in Ref. [7]. The variance of the x_i data, $\sigma^2(x_i)$, is calculated from Eq. (6a); and the errors in the mean are calculated from Eq. (8) when the Friedberg and Cameron method is used, $\sigma_{\mathrm{FC}}(\bar{x})$, and from Eq. (11) when the Straatsma *et al.* method is used, $\sigma_{\mathrm{S}}(\bar{x})$.

TABLE I

Results with Verlet and Toxvaerd Algorithms

N	$1+2\tau$	SI	$\sigma^2(x_i)\times 10^{-4}$	$\sigma_{\mathrm{FC}}(\bar{x})\times 10^{-3}$	$\sigma_{\mathrm{S}}(\bar{x})\times 10^{-3}$
170	5.04	4	1.00	1.5	1.7
143	5.71	6	0.81	1.8	1.8

Note. Verlet (first row) and Toxvaerd (second row) algorithms, where N is the total number of data points, τ is the correlation length, SI is the statistical inefficiency. The variance $\sigma^2(x_i)$ is calculated from Eq. (6a), and the errors in the means, σ_{FC} and σ_{S}, are calculated from Eqs. (8) and (11), respectively.

IV. Conclusions

The Friedberg and Cameron and the Straatsma *et al.* methods of estimating the statistical error in the data series have the advantage that they can be applied to data from any kind of simulation, MD or MC, whatever the constraints at constant temperature and/or pressure [8]. The time needed to obtain the correlations is negligible compared with the time taken to obtain the data from the simulation and it means that a big computer is not necessary to do these calculations.

We have seen mathematically how the statistical inefficiency and the correlation length are related for the general case. From the values in Table I, one can see that that relation is verified when the Toxvaerd algorithm is used, but when the Verlet algorithm is used the value of SI is less than $1 + 2\tau$ and the error in the mean is imprecise. Thus the accuracy of the algorithm in integrating the equations of motion for the particles is very important to obtaining the exact values of τ and SI, especially in the melting zone where the fluctuations are bigger. In our case, the Toxvaerd algorithm is more exact because it is a fifth-order predictor, while the Verlet algorithm is only a third-order predictor. It means that, with the Toxvaerd algorithm, the positions of the particles are very exactly calculated in the neighborhood of the repulsive part of the LJ potential, giving a higher value of the correlation in the data series and diminishing the uncertainty in the plotting respect to the results from the Verlet algorithm. This behavior can be seen through all the figures. Whereas the fluctuations in Fig. 2 are "regular," in Fig. 1 they are bigger and irregular and the statistical innefficiency obtained may be less precise. The analytical method, Ref. [4], is therefore better than the graphical method, Refs. [1–2], for obtaining the statistical inefficiency of the correlated data series.

Acknowledgment

This research was supported by the Comision Asesora Cientifica y Técnica (CAICYT), project No. 1071–84.

References

1. R. Friedberg and J. E. Cameron, *J. Chem. Phys.* **52**, 6049 (1970).
2. D. Fichman, N. Quirke, and D. J. Tildesley, *J. Chem. Phys.* **84**, 4535 (1986).
3. E. B. Smith and B. H. Wells, *Mol. Phys.* **53**, 701 (1984).
4. T. P. Straatsma, H. J. C. Berendsen, and A. J. Stam, *Mol. Phys.* **57**, 89 (1986).
5. L. Verlet, *Phys. Rev.* **159**, 98 (1967).
6. S. Toxvaerd, *J. Comput. Phys.* **47**, 444 (1982).
7. J. J. Morales, F. Cuadros, and L. F. Rull, *J. Chem. Phys.* **86**, 2960 (1987).
8. S. K. Schifel and D. C. Wallace, *J. Chem. Phys.* **83**, 5203 (1985).