JOURNAL OF COMPUTATIONAL PHYSICS 22, 319–329 (1976)

Efficient Analysis of Molecular Dynamics Data

T. R. KOEHLER

IBM Research Laboratory, San Jose, California 95193

AND

P. A. LEE

Bell Telephone Laboratories, Murray Hill, New Jersey 07974

Received February 9, 1976

The problem of obtaining the space–time Fourier transform of the displacement–displacement correlation function from the results of a molecular dynamics calculation is considered. It is found that the correlation function is most efficiently constructed if one selects only one data point from each statistically independent space–time element and that noise is suppressed in the Fourier transform if one averages over frequency and wavevector, which is equivalent to introducing damping factors in integrals over the space and time variables. Results of numerical experiments which support these conclusions are given and a comment pertaining to the adequacy of a molecular dynamics algorithm is made.

I. INTRODUCTION

The molecular dynamics MD method, the computer solution of Newton's equations of motion, was first applied by Alder and Wainright [1] to investigate the properties of a classical hard sphere system. The method has by now been extensively used [2–9]. The reason for its popularity is quite simple: It is the most direct technique that can be used to calculate time-dependent properties of an arbitrary, classical many-body system. In practice, of course, one can only simulate the behavior of a finite system during a finite time interval. Even with the most powerful modern computers, the limitations are on the order of a few thousand particles for ten to one hundred thousand time intervals of the maximum size permitted by considerations of numerical stability for the algorithm used to solve the differential equations.

After an MD run, one has available a collection of space–time data points and is faced with the task of obtaining as much statistically significant information from this data as possible. In this paper, a method for the efficient evaluation of the

displacement–displacement (u–u) correlation function and its space–time Fourier transform $S^1(q, \omega)$ will be described. The fundamental importance of this quantity has been emphasized by, for example, Martin [10] and is now generally appreciated. Two concepts are used in our method: (1) The most efficient way to accumulate data involves selecting only one data point from each statistically independent space–time volume. (2) Noise is suppressed in the Fourier transforms if one averages the time and space transforms over frequency and wavevector, respectively.

To date, we have only applied the data analysis technique to one model problem: a linear chain of anharmonic oscillators with a harmonic coupling between nearest neighbors. This problem has received considerable attention recently [11–16] for a variety of reasons which are irrelevant here. For our purposes, the attractive feature of the model is that the force law is quite simple so forces can be computed quickly and rather long MD runs are still economical.

In the next section we will describe the model and exhibit the usual expression for $S^1(q, \omega)$. In Section III, reasons for using an alternative expression will be given, the alternative will be derived and an efficient numerical procedure for calculating the u–u correlation function will be described. Finally, in Section IV, some numerical results will be presented and discussed. For convenience, we focus on the u–u correlation function only, although the ideas are trivially extended in whole to the velocity–velocity (v–v) correlation function and in part to the density-density correlation function.

II. The Problem

The interaction potential for the model system is

$$V(u_1, \ldots, u_N) = \sum_{n=1}^{N} [\tfrac{1}{4}(u_n^2 - 1)^2 + \tfrac{1}{2}c(u_{n+1} - u_n)^2], \tag{1}$$

where u_n is the displacement of the nth particle from its equilibrium position, there are N particles and periodic boundary conditions are used so that $u_{N+n} = u_n$. In this model, mass = 1 and time, displacement, temperature $T \equiv \langle v_n^2 \rangle$, and energy may be measured in appropriate units so that c and the dimensionless temperature T are the only parameters needed to cover the whole range of possible models of this general form.

In an MD calculation, one starts the system with initial displacements and velocities which are designed to be fairly representative for the desired temperature. An algorithm is used which generates the positions at a time $t + h$ from those at t. The Verlet [3] central difference algorithm

$$u_n(t + h) = 2u_n(t) - u_n(t - h) + h^2\ddot{u}_n(t) \tag{2}$$

with $h = 0.2$ was found to be convenient and adequate. One starts the calculation, allows the system to run until any nonrandom initial correlations have died out, calls this the zero of time and collects the desired dynamical information from that point on. However, we are not concerned here with the technical details of the MD method. Such may be found in [1–9]. In the remainder of this paper, it will simply be assumed that the displacements $u_n(jh)$ are known for $n = 1, N$ at the times jh with $j = 1, M$.

The quantity S^1 is defined by

$$S^1(q, \omega) = \int_{-\infty}^{\infty} dt \sum_n e^{-i\omega t} e^{iqn} \langle u_n(t)\, u_0(0) \rangle, \tag{3}$$

where the angular brackets denote an ensemble average, which is equal to the time average for an ergodic system. The wavevector q is defined in the usual way as $q = 2\pi l/N$, $-N/2 < l \leqslant N/2$. The symbol S^1 is used because this quantity is the leading approximation to the dynamic structure factor $S(q, \omega)$, the Fourier transform of the density–density correlation function. If the ensemble average is replaced by an average over all space and time data points obtained in MD, one obtains

$$\langle u_i(nh)\, u_0(0) \rangle = (1/NM) \sum_{m=1}^{M} \sum_{j=1}^{N} u_{i+j}(mh + nh)\, u_j(mh), \tag{4}$$

where, in space, cyclic boundary conditions have been used and, in time, it is assumed that M is sufficiently greater than the interesting range of n so that end effects may be ignored. When Eq. (4) is inserted into Eq. (3), it is easy to show that

$$S^1(q, \omega) = (1/M)\, |\, \rho_q(\omega)|^2, \tag{5}$$

where

$$\rho_q(\omega) = (1/2\pi) \sum_{j=1}^{M} e^{i\omega jh} X_q(jh) \tag{6}$$

and

$$X_q(t) = (1/N^{1/2}) \sum_{n=1}^{N} e^{inq} u_n(t). \tag{7}$$

Equations (5)–(7) are rather convenient for the calculation of $S^1(q, \omega)$ for one particular value of q, since $X_q(t)$ can be evaluated at each time step and no reuse or even saving of intermediate data is required.

A plot of $S^1(0, \omega)$ obtained from the straightforward application of Eqs. (5)–(7) for runs using between 500 and 10,000 time steps are shown in Figs. 1a–1d. All are for 1000 particles. Although their values are irrelevant for our purposes here, the temperature was $T = 1.082$ and the nearest neighbor force constant was

$c = 0.25$. These three values and the choice $q = 0$ will be the only cases treated here. It is obvious that there is a problem associated with the graphs of Fig. 1: The results appear to be mostly noise, and the noise increases as the time span increases although one would expect this to result in better statistical reliablility. Such noise is also apparent in the plots of [6].

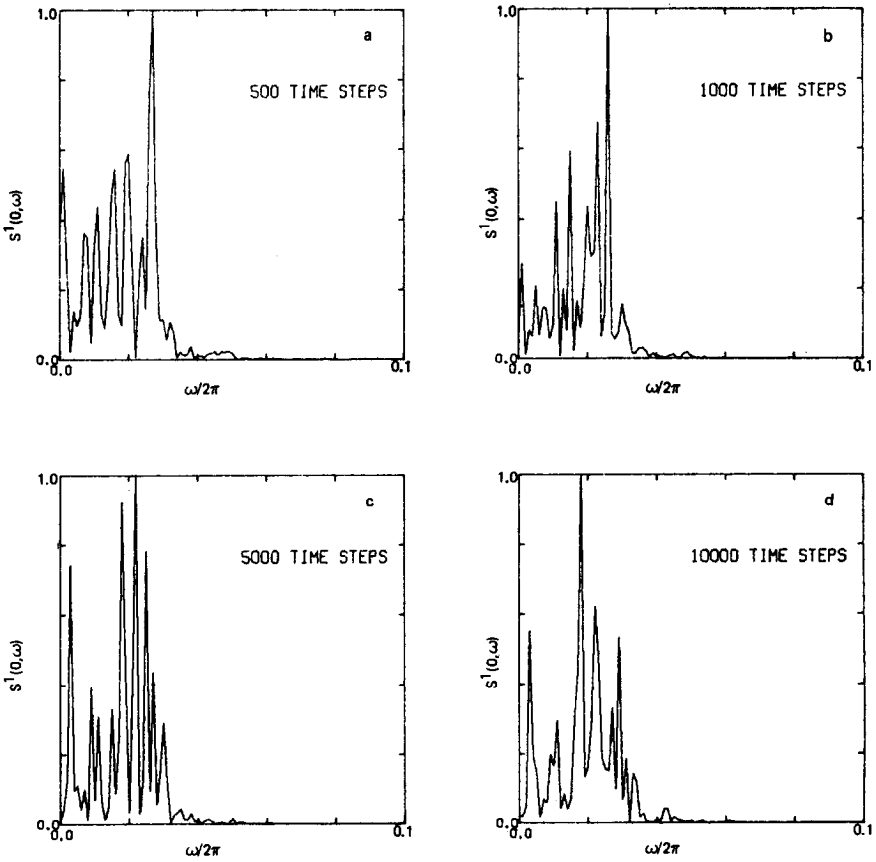

FIG. 1. $S^1(0, \omega)$ calculated according to Eqs. (5)–(7) for four different total time steps between 500 and 10,000 as indicated on the plots. The intensity is normalized so that the maximum is unity.

The origin of the problem is that (5) is an expression for S^1 at precisely one frequency and one wavevector. Using this expression is analogous to using a very narrow slit width in a high resolution spectroscopy experiment. In either case, the signal will be masked by noise. What has been done is to average over

neighboring frequencies [6, 9] or wavevectors [5] to produce a smoother curve. If a Gaussian averaging is used, we can formally define an averaged quantity

$$\bar{S}^1(q, \omega) = (2/\pi)(\Delta_q \Delta_\omega)^{1/2} \int_{-\infty}^{\infty} dq' \int_{-\infty}^{\infty} d\omega' \, S^1(q - q', \omega - \omega') \times \exp[-(q - q')^2/\Delta_q^2] \exp[-(\omega - \omega')^2/\Delta_\omega^2]. \tag{8}$$

It is easy to show that

$$\bar{S}^1(q, \omega) = \sum_n \sum_j C_j(nh) \cos(qj) \cos(\omega nh) \exp[-(\Delta_\omega nh/2)^2] \exp[-(\Delta_q j/2)^2], \tag{9}$$

where

$$C_j(t) = \langle u_j(t) \, u_0(0) \rangle \tag{10}$$

has been used to represent the u–u correlation functions and the reflection symmetry of $C_j(t)$ about $j = 0$ and $t = 0$ has been used.

Thus we see that the concept of using Gaussian averages over frequency and wavevector formally leads to the introduction of Gaussian damping factors in space and time. These factors effectively set correlations between points which are widely separated in space or time equal to zero. Consideration of the correlation function in space and time provides an alternate viewpoint for the source of the high noise level associated with the use of (5)–(7). The following discussion of this viewpoint will be intuitive rather than rigorous, but will later be supplemented by supporting evidence from the results of numerical experiments.

As an illustration, we have computed the correlation function $\langle X_{q=0}(t) \, X_{q=0}(0) \rangle$ by summing over all particles according to Eq. (7) at each time step and then averaging over every time step according to Eq. (4). The result is shown in Fig. 2a. Clearly, everything beyond about 150 time steps is noise. Although it is not shown here, the noise level remains essentially constant if the time span of the plot is extended, which explains the severe noise problem exhibited in Fig. 1 even for very long runs. The time Fourier transform of $\langle X_q(t) \, X_q(0) \rangle$ over the entire range of a run is mathematically equivalent to using Eqs. (5)–(7). However, in the former procedure, it is clear that only the first 150 or so time steps will contribute physically meaningful structure to S^1, the remainder will all be noise. We have checked that, if the transform of $\langle X_0(t) \, X_0(0) \rangle$ be carried out for the entire time span of the run, S^1 as shown in Fig. 1 is recovered.

It is now clear why running the MD for a longer period of time T did not make Fig. 1b an improvement over Fig. 1a. While the noise level will be reduced like $T^{1/2}$, the amount of noisy data used in the Fourier transform increases like T. A Gaussian damping in time, or equivalently, a Gaussian averaging in frequency, suppresses the noisy part of the correlation function and yields physically meaningul results. Similar considerations apply to spatial correlations and wavevector averaging.

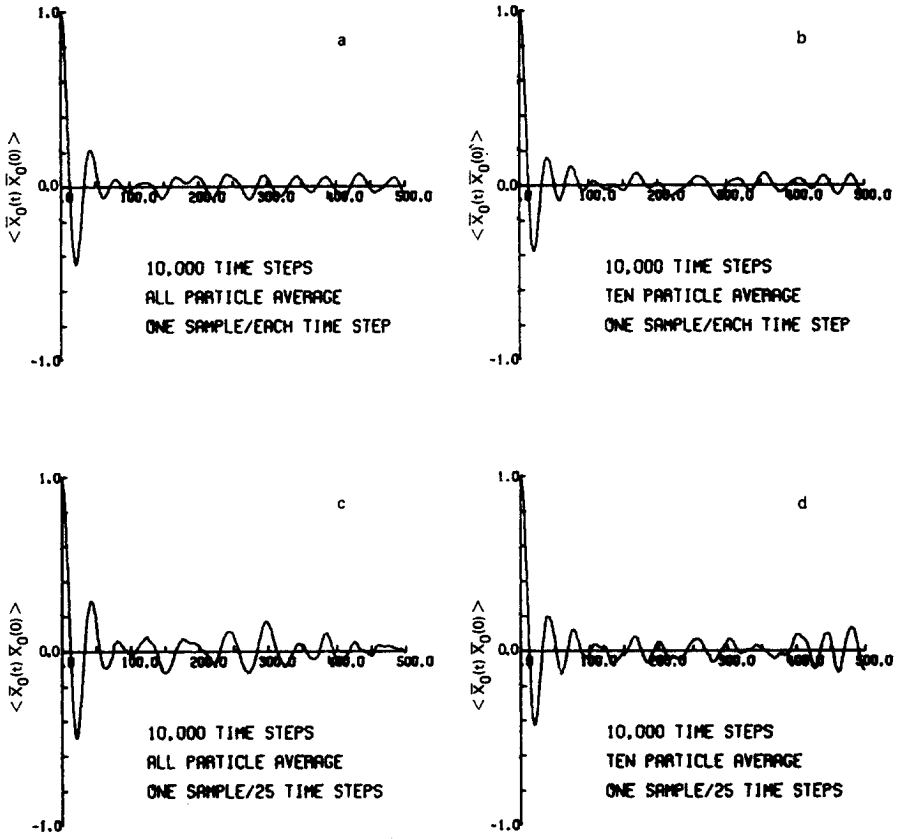

FIG. 2. $\langle \bar{X}_0(t)\bar{X}_0(0)\rangle$ from a 10,000 time step MD run calculated by various sampling procedures as indicated on the plots, where $\bar{X}_0(t) = (1/N)\Sigma u_n(t)$. The curves are normalized to unity at $t = 0$.

III. EFFICIENT DATA ANALYSIS

Computation according to Eqs. (5)–(7) and then frequency and wavevector averaging requires storage of a large number of q and ω values to be averaged at the end of the computation. Indeed, the ultimate accuracy is determined by the square root of the number of stored $S^1(q + q_1, \omega + \omega_1)$ values for each value of q and ω. Thus it is desirable to compute C_j which contains all the information required to obtain any $\bar{S}^1$ according to Eq. (9). However, the direct computation of the correlation function according to Eq. (4) requires a large number of operations and storage. What we will discuss next is a procedure for computing

$C_j(nh)$ which makes optimal use of the satistically independent parts of the data. The method has the advantage that the amount of storage required is not determined by the desired accuracy, but rather by the spatial and temporal coherence lengths.

It is well known [11, 13, 15] that spatial coherence for the model system (1) extends over only a few lattice sites, except at temperatures which are well below that considered here. Thus particles separated by more than this are nearly uncorrelated and, in a chain of length 1000, there are perhaps 100 essentially independent segments of length 10. In looking at only one q value (i.e., using an undamped 1000 particle sum to evaluate $X_q(t)$), one rather than 100 statistically independent samples are used. Similar considerations apply to the time Fourier transform. Configurations separated by more than a few correlation times are statistically independent samples; combining them into one estimate of $X_q(t)$ according to (7) does not take advantage of this.

The proper treatment of statistically independent space–time regions is also important for the efficient computation of $C_j(t)$, which, in the MD context, is done according to

$$C_j(nh) = (1/N_k M_m) \sum_{\{k\}} \sum_{\{m\}} u_{j+k}(nh + mh)\, u_k(mh), \tag{11}$$

where $\{k\}$ and $\{m\}$ indicate that the summation is over a selected set of N_k and M_m, k and m values, respectively. Considerations for the optimal selection of the contents of the sets will be given in the following.

It is clear that, if all available k and m values are used, this construction requires on the order of $N \times M$ arithmetic operations per (jn) pair, although the exact number decreases with increasing n. It requires on the order of the square of this quantity for all (jn) pairs. For the reasonable values $N = 10^3$, $M = 10^4$ this is on the order of 10^{14} arithmetic operations, an obviously prohibitive amount. This requirement can be cut down enormously, however. First of all, j and n need only span a few space N_j and time N_n coherence lengths, respectively, which is about $N_j \times N_n = 10 \times 300$ in the example used so far. This would still leave one with the unrealistic requirement of 3×10^{11} operations except for a second important consideration: Only contributions from (km) pairs which are uncorrelated give statistically independent contributions to the sum. Thus the k and m values in $\{k\}$ and $\{m\}$ should span the entire $N \times M$ range, but be separated by approximately one coherence length in space or time, respectively. This reduces the number of operations to a manageable size; furthermore, each operation makes a statistically significant contribution to the total so that any additional reduction would waste some of the original MD data.

The storage requirements for constructing C_j by this method are modest; one can do the calculations as the system evolves and does not have to save all

the $u_j(t)$. Two arrays need to be stored. One is $N_j \times N_n$ for the C_j elements. The other is N/(the number of elements in $\{k\}$ by N_n/(the number of elements in $\{m\}$) and contains the past $u_j(t)$ values which are still needed to obtain the current contributions to C_j.

The effect of different sampling methods is illustrated in Fig. 2 where plots of $\langle X_0(t)\, X_0(0)\rangle$ obtained in four different ways are shown. The raw data for each

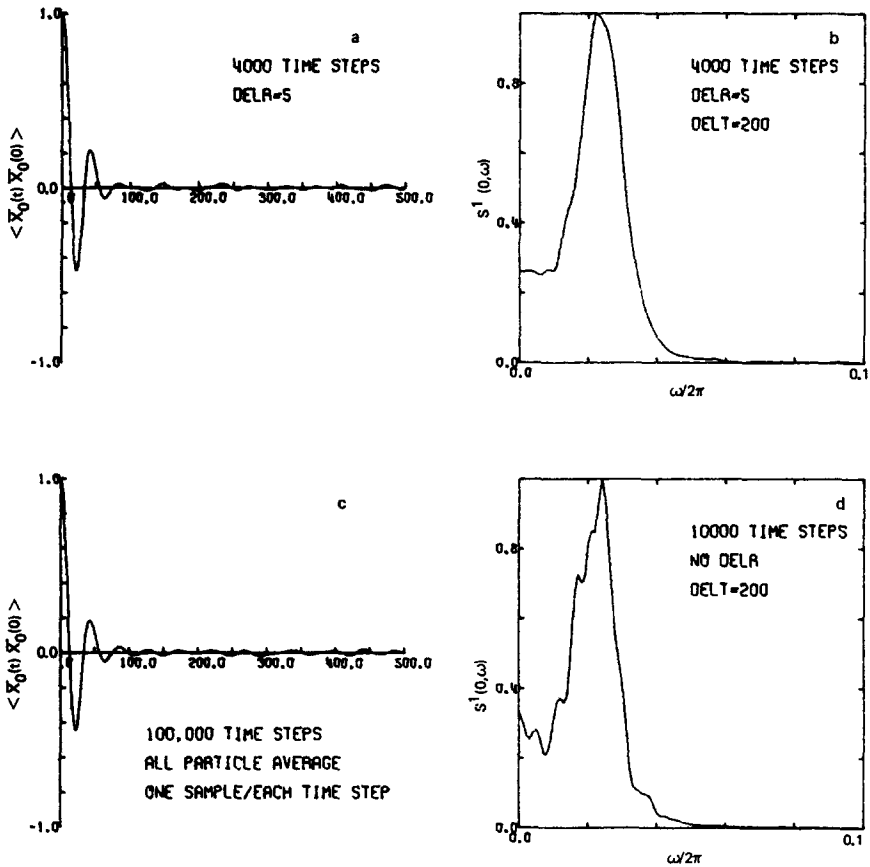

FIG. 3. (a) $\langle \bar{X}_0(t)\bar{X}_0(0)\rangle$ from a 4000 time step MD run calculated using spatial Gaussian damping with $1/e$ decay length of five lattice sites. The u–u correlation function was constructed from the raw data by sampling every fourth site in space and every 25th in time as an independent origin. (b) $S^1(0, \omega)$, the Fourier transform of the curve of Fig. 3a with its long time noise suppressed by a Gaussian damping factor with a $1/e$ decay time of 200 steps. The intensity is normalized so that the maximum is unity. (c) Similar to Fig. 2a except the MD run is extended to 100,000 time steps. (d) Related to the curve of Fig. 2a in the same way Fig. 3b is related to the curve of Fig. 3a.

plot was the same 10^4 time steps of the run used to obtain Fig. 1d. The displacements of all the particles in the chain were used to compute $X_0(t)$ in Figs. 2a and 2c, while the mean position of ten arbitrary adjacent particles was used in Figs. 2b and 2d. Each time step was used in m for Figs. 2a and 2b while $m = 1, 26, 51,...$ in Figs. 2c and 2d. In these plots, everything beyond about 150 time steps is noise. This is substantiated by comparison with additional studies using $M = 10^5$ and the smoothed plot of Fig. 3a to be discussed later. It is clear that there is very little difference in the noise level in the four plots of Fig. 2 even though each point in the plot of Fig. 2a has 2500 times as many contributions as in Fig. 2d. This substantiates the claim that the noise level can only be suppressed by the accumulation of statistically independent data.

Finally, the effects of the smoothing procedure are shown in Fig. 3. An intermediate quantity $\langle \bar{X}_0(t)\, \bar{X}_0(0) \rangle$ is shown in Fig. 3a, where the averaging symbols denote averaging with respect to wavevector only. Wavevector averaging is particularly attractive for this model since the spatial coherence length [11, 13, 15] is known exactly. The contrast with the noise level of Fig. 2 is striking and compares very favorably to that of Fig. 3c, which was obtained by extending the run for Fig. 2a to 10^5 time steps. For the wavevector smoothing $(\Delta_q/2) = 1/5$ and, in the construction of $C_j(t)$, $\{k\} = 1, 5, 9,..., 997$ was used. The frequency averaged, with $(\Delta_\omega/2) = 200$, time Fourier transform of the curve of Fig. 3a is shown in Fig. 3b. The contrast with Fig. 1 is more than striking. A similar plot starting with the curve of Fig. 2a is shown in Fig. 3d. The noise in this plot could be suppressed more by increasing Δ_ω, but it is still a considerable improvement over Fig. 1. The $S^1(0, \omega)$ curve obtained from Fig. 3c is not shown here, but it is very similar to Fig. 3b.

IV. Summary

We have shown the desirability of computing correlation functions during the analysis of MD data and the importance of using statistically independent quantities in this computation. First, averaging Fourier transforms over wavevectors or frequency introduced damping factors in space or time which separated statistically uncorrelated data and this averaging can be most efficiently implemented if one starts with the correlation function. Second, selection of only a few points from each statistically independent space–time volume minimized the computational effort required to construct the correlation function without diminishing the signal to noise ratio.

As is obvious from the references, many of the observations made in this paper have appeared elsewhere. The correlation function in [8] was constructed using selected initial points for m as in Eq. (11). $S^1(q, \omega)$ is simply the power spectrum

of X_q, a familiar [17, 18] practical application of Fourier transforms. The relation between frequency averaging and time damping follows from the convolution theorem as was noted in [9]. Thus the simple cutoff in time at T_m used in [4] and [7] is equivalent to averaging in frequency with $\sin[(\omega - \omega') T_m]/(\omega - \omega')$. Such filtering is extensively discussed in works such as [17] and [18], although the specific case of Gaussian averaging is rarely treated, probably because it is difficult to accomplish with electrical or electronic circuitry.

We feel that our major contribution here has been to point out various consequences of statistical independence, to note their interconnections and their implications for MD data analysis, to recommend numerical procedures for the latter and to provide the results of numerical experiments which tested the recommendations. In addition, while the connection between frequency averaging and damping in time was noted in [9] and is implicit in [17, 18] and similar works, we have not seen a previous application of the similar relationship between wavevector averaging and spatial damping.

In addition, we would like to emphasize a point of view, which has also been expressed by Hoover and Ashurst [19], regarding criteria for the adequacy of an MD algorithm: It is not necessary that the algorithm solve the initial value problem accurately over the entire length of the run. Rather, the solutions must conserve energy throughout the entire run, so that one is looking at a unique ensemble, but need only be accurate for a time on the order of one temporal coherence length, because data separated by longer times are essentially uncorrelated. The presence of longer term inaccuracies will be roughly equivalent to averaging over a series of short runs with a proper statistical distribution of initial conditions. From this viewpoint, we found that the Verlet [3] algorithm had quite good long an short term energy conservation, although to see the latter we had to use a higher order interpolation formula

$$h\dot{u}_n(t) = (1/2)[u_n(t + h) - u_n(t - h)] - (h^2/12)[\ddot{u}_n(t + h) - \ddot{u}_n(t - h)] \quad (12)$$

for the velocity than was used in [3].

The application of the concepts of this paper to the construction of the density–density correlation function is straightforward, but the numerical implementation of the resulting expressions is not as obvious and may prove to be difficult in practice. We have not attempted to explore applications to correlation functions in liquids.

REFERENCES

1. B. J. ALDER AND T. E. WAINWRIGHT, *J. Chem. Phys.* **33** (1960), 1439.
2. A. RAHMAN, *Phys. Rev.* **136** (1964), A405.
3. L. VERLET, *Phys. Rev.* **159** (1967), 98.

4. A. RAHMAN AND F. H. STILLINGER, *J. Chem. Phys.* **55** (1971), 3336.
5. D. LEVESQUE, L. VERLET, AND J. KÜRKIJARVI, *Phys. Rev.* **A7** (1973), 1690.
6. J. P. HANSEN AND M. L. KLEIN, *J. Phys. (Paris)* **35** (1974), L–29.
7. F. STILLINGER AND A. RAHMAN, *J. Chem. Phys.* **60** (1974), 1545.
8. D. LEVESQUE AND W. T. ASHURST, *Phys. Rev. Lett.* **33** (1974), 277.
9. T. SCHNEIDER AND E. STOLL, *Phys. Rev.* **B13** (1976), 1216.
10. P. C. MARTIN, *in* "1967 Les Houches Lectures," p. 39. Gordon and Breach, New York, 1968.
11. D. J. SCALAPINO, M. SEARS, AND R. A. FERREL, *Phys. Rev.* **B6** (1972), 3409.
12. K. K. MURATA, *Phys. Rev.* **B11** (1975), 462.
13. J. A. KRUMHANSL AND J. R. SCHRIEFFER, *Phys. Rev.* **B11** (1975), 3535.
14. T. SCHNEIDER AND E. STOLL, *Phys. Rev. Lett.* **35** (1975), 296.
15. S. AUBRY, *J. Chem. Phys.* **62** (1975), 3217.
16. T. R. KOEHLER, A. R. BISHOP, J. A. KRUMHANSL, AND J. R. SCHRIEFFER, *Solid State Commun.*, **17** (1975), 1515.
17. D. C. CHAMPENEY, "Fourier Transforms and Their Physical Applications," Academic Press, New York, 1973.
18. R. B. BLACKMAN AND J. W. TUCKEY, "The Measurement of Power Spectra," New York, 1958.
19. W. G. HOOVER AND W. T. ASHURST, *in* "Theoretical Chemistry Advances and Perspectives," Vol. 1, p. 1. Academic Press, New York, 1975.