Appendix D

Statistical Errors

It is often stated that a computer simulation generates "exact" data for a given model. However, this is true only if we can perform an infinitely long simulation. In practice we have neither the budget nor the patience to approximate such a simulation. Therefore, the results of a simulation will always be subjected to statistical errors, which have to be estimated.

D.1 Static Properties: System Size

Let us consider the statistical accuracy of the measurement of a dynamical quantity A in a Molecular Dynamics simulation (the present discussion applies, with minor modifications, to Monte Carlo simulations). During a simulation of total length T, we obtain the following estimate for the equilibrium-average of A:

$$A_\tau = \frac{1}{\tau}\int_0^\tau dt\, A(t), \qquad \text{(D.1.1)}$$

where the subscript on A_τ refers to averaging over a finite "time" τ. If the ergodic hypothesis is justified then $A_\tau \to \langle A \rangle$, as $\tau \to \infty$, where $\langle A \rangle$ denotes the ensemble average of A. Let us now estimate the variance in A_τ, $\sigma^2(A)$:

$$\begin{aligned} \sigma^2(A) &= \langle A_\tau^2 \rangle - \langle A_\tau \rangle^2 \\ &= \frac{1}{\tau^2}\int_0^\tau\int_0^\tau dtdt'\ \langle [A(t) - \langle A \rangle]\,[A(t') - \langle A \rangle] \rangle\,. \qquad \text{(D.1.2)} \end{aligned}$$

Note that $\langle [A(t) - \langle A \rangle][A(t') - \langle A \rangle] \rangle$ in equation (D.1.2) is simply the time correlation function of fluctuations in the variable A. Let us denote this correlation function by $C_A(t - t')$. If the duration of the sampling τ is much larger than the characteristic decay time t_A^c of C_A, then we may rewrite equation (D.1.2) as

$$\begin{aligned}\sigma^2(A) &\approx \frac{1}{\tau}\int_{-\infty}^{\infty} dt\, C_A(t)\\ &\approx \frac{2t_A^c}{\tau}C_A(0).\end{aligned} \qquad (D.1.3)$$

In the last equation we have used the definition of t_A^c as half the integral from $-\infty$ to ∞ of the normalized correlation function $C_A(t)/C_A(0)$. The relative variance in A_τ therefore is given by

$$\frac{\sigma^2(A)}{\langle A\rangle^2} \approx (2t_A^c/\tau)\frac{\langle A^2\rangle - \langle A\rangle^2}{\langle A\rangle^2}. \qquad (D.1.4)$$

Equation (D.1.4) clearly shows that the root-mean-square error in A_τ is proportional to $\sqrt{t_A^c/\tau}$. This result is hardly surprising. It simply states the well-known fact the variance in a measured quantity is inversely proportional to the number of uncorrelated measurements. In the present case, this number is clearly proportional to τ/t_A^c. This result may appear to be trivial, but it is nevertheless very important, because it shows directly how the lifetime and amplitude of fluctuations in an observable A affect the statistical accuracy. This is of particular importance in the study of fluctuations associated with hydrodynamical modes or pretransitional fluctuations near a symmetry-breaking phase transition. Such modes usually have a characteristic lifetime proportional to the square of their wavelengths. To minimize the effects of the finite system size on such phase transitions, it is preferable to study systems with a box size L large compared with all relevant correlation lengths in the system. However, due to the slow decay of long-wavelength fluctuations, the length of the simulation needed to keep the relative error fixed should be proportional to L^2. As the CPU time for a run of fixed length is proportional to the number of particles (at best), the CPU time needed to maintain constant accuracy increases quite rapidly with the linear dimensions of the system (e.g., as L^5 in three dimensions).

Another aspect of equation (D.1.4) is not immediately obvious; namely, it makes a difference whether the observable A can be written as a sum of uncorrelated single-particle properties. If this is the case, then it is easy to see that the ratio $(\langle A^2\rangle - \langle A\rangle^2)/\langle A\rangle^2$ is inversely proportional to the number of particles, N. To see this, consider the expressions for $\langle A\rangle$ and $\langle A^2\rangle - \langle A\rangle^2$ in this case:

$$\langle A\rangle = \sum_{i=1}^{N}\langle a_i\rangle = N\langle a\rangle \qquad (D.1.5)$$

and

$$\langle A^2\rangle - \langle A\rangle^2 = \sum_{i=1}^{N}\sum_{j=1}^{N}\langle[a_i - \langle a\rangle][a_j - \langle a\rangle]\rangle. \qquad (D.1.6)$$

If the fluctuations in a_i and a_j are uncorrelated, then we find that

$$\frac{\langle A^2\rangle - \langle A\rangle^2}{\langle A\rangle^2} = \frac{1}{N}\frac{\langle a^2\rangle - \langle a\rangle^2}{\langle a\rangle^2}. \qquad (D.1.7)$$

From equation (D.1.7) it is clear that the statistical error in a single-particle property is inversely proportional to $\sqrt{N}$. Hence, for single-particle properties, the accuracy improves as we go to larger systems (at fixed length of the simulation). In contrast, no such advantage is to be gained when computing truly collective properties.

D.2 Correlation Functions

We can apply essentially the same arguments to estimate the statistical errors in time correlation functions. Suppose that we wish to measure the (auto)correlation function[1] of the dynamical quantity A. To obtain an estimate of $C_A(\tau) \equiv \langle A(0)A(\tau)\rangle$, we average the product $A(t)A(t+\tau)$ over the initial time t. Suppose that the length of our run is τ_0, then our estimate of $C_A(\tau)$ is

$$\overline{C}_A(\tau) = 1/\tau_0\int_0^{\tau_0} dt\, A(t)A(t+\tau),$$

where the bar over C_A denotes the average over a finite time τ_0. Next, we consider the variance in $\overline{C}_A(\tau)$ [527]:

$$\begin{aligned}&\langle\overline{C}_A(\tau)^2\rangle - \langle\overline{C}_A(\tau)\rangle^2\\ &= (1/\tau_0{}^2)\int_0^{\tau_0}\int_0^{\tau_0} dt'dt''\ \langle A(t')A(t'+\tau)A(t'')A(t''+\tau)\rangle\\ &\quad -(1/\tau_0{}^2)\int_0^{\tau_0}\int_0^{\tau_0} dt'dt''\ \langle A(t')A(t'+\tau)\rangle\langle A(t'')A(t''+\tau)\rangle.\end{aligned} \qquad (D.2.1)$$

The first term on the right-hand side of equation (D.2.1) contains a fourth-order correlation function. To simplify matters, we shall assume that the fluctuations of A follow a Gaussian distribution. This is not the simple Gaussian distribution that describes, for instance, the Maxwell distribution

[1] The extension to cross-correlation functions of the type $\langle A(t)B(0)\rangle$ is straightforward and left as an exercise to the reader.

of particle velocities in equilibrium, but a multidimensional (in fact, infinite-dimensional) distribution that describes all correlations between fluctuations of A at different times. For the simple case that we consider only real fluctuations at discrete times, this distribution would be of the following form:

$$P(A(t_1), A(t_2), \cdots, A(\tau_{0n})) = \text{const.} \times \exp\left[-\frac{1}{2}\sum_{i,j} A(t_i)\alpha(t_i - t_j)A(t_j)\right],$$

where the matrix $\alpha(t_i - t_j)$ is simply the inverse of the (discrete) time correlation function $C_A(t_i - t_j)$. For Gaussian variables, we can factorize all higher-order correlation functions. In particular,

$$\begin{aligned} &\langle A(t')A(t'+\tau)A(t'')A(t''+\tau)\rangle \\ &= \langle A(t')A(t'+\tau)\rangle \langle A(t'')A(t''+\tau)\rangle \\ &+ \langle A(t')A(t'')\rangle \langle A(t'+\tau)A(t''+\tau)\rangle \\ &+ \langle A(t')A(t''+\tau)\rangle \langle A(t'+\tau)A(t')\rangle . \end{aligned} \tag{D.2.2}$$

Inserting equation (D.2.2) in equation (D.2.1), we get

$$\begin{aligned} &\langle \overline{C}_A(\tau)^2\rangle - \langle \overline{C}_A(\tau)\rangle^2 \\ &= (1/\tau_0{}^2)\int_0^{\tau_0}\int_0^{\tau_0} dt'dt'' \; \langle A(t')A(t'')\rangle \langle A(t'+\tau)A(t''+\tau)\rangle \\ &+ (1/\tau_0{}^2)\int_0^{\tau_0}\int_0^{\tau_0} dt'dt'' \; \langle A(t')A(t''+\tau)\rangle \langle A(t'+\tau)A(t'')\rangle \\ &= (1/\tau_0{}^2)\int_0^{\tau_0}\int_0^{\tau_0} dt'dt'' \; \langle A(t'-t'')A(0)\rangle^2 \\ &+ (1/\tau_0{}^2)\int_0^{\tau_0}\int_0^{\tau_0} dt'dt'' \; \langle A(t'-t''-\tau)A(0)\rangle \langle A(t'-t''+\tau)A(0)\rangle . \end{aligned} \tag{D.2.3}$$

Again, we consider the case where the length of the simulation, τ_0, is much longer than the characteristic decay time of the fluctuations of A. In that case, we can write

$$\begin{aligned} &\langle \overline{C}_A(\tau)^2\rangle - \langle \overline{C}_A(\tau)\rangle^2 \\ &= (1/\tau_0)\int_{-\infty}^{\infty} dx \left(\langle A(x)A(0)\rangle^2 + \langle A(x-\tau)A(0)\rangle \langle A(x+\tau)A(0)\rangle\right), \end{aligned} \tag{D.2.4}$$

where we have defined the variable x as $t' - t''$. Let us now consider two limiting cases, $\tau = 0$ and $\tau \to \infty$. For $\tau = 0$, we can write

$$\begin{aligned} \langle \overline{C}_A(\tau)^2\rangle - \langle \overline{C}_A(\tau)\rangle^2 &= (2/\tau_0)\int_{-\infty}^{\infty} dx\, \langle A(x)A(0)\rangle^2 \\ &= 4\left\langle A^2(0)\right\rangle^2 \frac{\tau^c}{\tau_0}. \end{aligned} \tag{D.2.5}$$

The last line of this equation defines the correlation time τ^c:

$$\tau^c \equiv \frac{\int_0^\infty dx\, \langle A(x)A(0)\rangle^2}{\langle A^2(0)\rangle^2}.$$

For $\tau \to \infty$, the product

$$\langle A(x-\tau)A(0)\rangle \langle A(x+\tau)A(0)\rangle$$

vanishes, and we have

$$\langle \overline{C}_A(\tau)^2\rangle - \langle \overline{C}_A(\tau)\rangle^2 = 2\left\langle A^2(0)\right\rangle^2 \frac{\tau^c}{\tau_0}. \tag{D.2.6}$$

Comparison of equation (D.2.5) with equation (D.2.6) shows that the *absolute* error in $C_A(\tau)$ changes only little with τ. As a consequence, the *relative* error in time correlation functions increases rapidly as $C_A(\tau)$ decays to 0. In this derivation we have assumed that the total number of samples for each τ is equal; in case we have (many) fewer samples for large τ, this approach is not valid. In fact, if we have many fewer samples for large values of τ, we may wonder whether these values are reliable.

It should be stressed that the preceding error estimate is only approximate, because it relies on the validity of the Gaussian approximation. In specific cases (e.g., for the fluctuations of particle velocities), it is known that deviations from the Gaussian approximation occur. However, the deviations (where they are known) are usually not large, and it seems likely that error estimates based on the Gaussian approximation are on the correct order of magnitude. Very little evidence, however, supports or contradicts this belief.

A more detailed discussion of statistical errors in collective and single-particle time correlation functions can be found in [527] and [528]. Systematical techniques for measuring statistical errors in a simulation are discussed in [529] and [19].

D.3 Block Averages

The previous section showed that we can estimate the statistical error in time correlation functions on the basis of our knowledge of the time correlation function itself. Hence, no extra work is needed to arrive at an error estimate. However, as discussed in section D.1, to arrive at an error estimate for a static quantity, we need to compute the time correlation function of that quantity. As the computational effort to compute a time correlation function is larger than that required for a static average, we usually estimate statistical errors in static quantities by studying the behavior of so-called block averages. A block average is simply a time average over a finite time t_B:

$$\bar{A}_B \equiv \frac{1}{t_B}\int_0^{t_B} dt\, A(t).$$

During a simulation, we can easily accumulate block averages for a given block length t_B. After the simulation has been completed, we can compute the block averages for blocks of length $n \times t_B$ by simply averaging the block averages of n adjacent blocks of length t_B. Let us now consider the variance in the block averages for a given value of t_B:

$$\sigma^2(\bar{A}_B) = \frac{1}{n_B}\sum_{b=1}^{n_B} \left(\bar{A}_B - \langle A \rangle\right)^2.$$

If t_B is much larger than the correlation time t_A^c, we know from section D.1 that

$$\sigma^2(\bar{A}_B) \approx \left(\langle A^2 \rangle - \langle A \rangle^2\right)\frac{t_A^c}{t_B}. \qquad \text{(D.3.1)}$$

But, as yet, we do not know t_A^c. We therefore compute the product

$$P(t_B) \equiv t_B \times \frac{\sigma^2(\bar{A}_B)}{\langle A^2 \rangle - \langle A \rangle^2}.$$

In the limit $t_B \gg t_A^c$, we know that $P(t_B)$ must approach t_A^c. Hence, we plot $P(t_B)$ versus t_B (or, more conveniently, $1/P(t_B)$ versus $1/t_B$) and estimate the limit of $P(t_B)$ for $t_B \to \infty$. This yields our estimate of t_A^c and thereby our error estimate for $\bar{A}$. This analysis of block averages is a very powerful tool to determine whether a simulation is long enough to yield a reliable estimate of a particular quantity: if we find that $P(t_B)$ is still strongly dependent on t_B in the limit $t_B = \tau$, then we know that our run is too short.

An alternative method for estimating the statistical error in a simulation has been developed by Flyvbjerg and Petersen [84]. Let $A_1, A_2, \ldots, A_L$ be L consecutive samples of some fluctuating quantity A of which we want to calculate its ensemble average and statistical error. We assume that all L samples are taken after the system has been equilibrated. The ensemble average is estimated from

$$\langle A \rangle \approx \bar{A} \equiv \frac{1}{L}\sum_{i=1}^{L} A_i, \qquad \text{(D.3.2)}$$

and we need an estimator of the variance of

$$\sigma^2(A) = \langle A^2 \rangle - \langle A \rangle^2 \approx \frac{1}{L}\sum_{i=1}^{L} \left[A_i - \bar{A}\right]^2. \qquad \text{(D.3.3)}$$

If all L samples were uncorrelated, we could use the standard formulas of statistics to calculate this variance. However, in a simulation, the samples are correlated and we have to take this correlation into account.

The idea behind the method of Flyvbjerg and Petersen is to group the simulation data into consecutive blocks and compute an average for each of these blocks. These block averages will show less and less correlation between two consecutive blocks if the block size is made larger. In the limit that there is no detectable correlation between the blocks, the standard statistical formulas are valid and the standard deviations as a function of the block size follow these formulas. This procedure leads to a reliable estimate of the standard deviation.

To see how this method works in practice, consider the following transformation of our data set $A_1, A_2, \ldots, A_L$ into a new data set $A'_1, A'_2, \ldots, A'_{L'}$, which has half the size of the original set:

$$A'_i = 0.5(A_{2i-1} + A_{2i})$$

with

$$L' = 0.5L.$$

Note that the average of the new set $\bar{A}'$ is the same as for the original one. The variance in $\bar{A}'$ is given by

$$\sigma^2(A') = \left\langle A'^2 \right\rangle - \langle A' \rangle^2 = \frac{1}{L'}\sum_{i=1}^{L'} A'^2_i - \bar{A}'^2.$$

We can continue to perform this blocking operation, and if we have performed the simulation sufficiently long, the averages A'_i will become completely uncorrelated. If this is the case the following relation should hold:

$$\frac{\sigma^2(A')}{L'-1} \approx \text{Constant}.$$

This constant value is used as an estimate of the variance in the ensemble average. Note that, in a similar way, we can also determine the statistical error in $\sigma^2(A')$. This gives as estimate of the variance in our ensemble average

$$\sigma^2(A) \approx \frac{\sigma^2(A')}{L'-1} \pm \sqrt{\frac{2\sigma^4(A')}{(L'-1)^3}}. \qquad \text{(D.3.4}$$

In Case Study 4, this method was used to compute the standard deviation of the energy in a Molecular Dynamics simulation. In Figure 4.4 a typical plot of this estimate of the variance as a function of block size is shown. For small values of M, the number of blocking operations, the data are correlated and as a consequence the variance will increase if we perform the blocking operation. For very high values of M we have only a few samples, and as a result, the statistical error in our estimate of $\sigma^2(A)$ will be large. The plateau in between gives us the value of $\sigma^2(A)$ we are interested in.

Appendix E

Integration Schemes

E.1 Higher-Order Schemes

The basic idea behind the predictor-corrector algorithms is to use information about the position and its first n derivatives at time t to arrive at a prediction for the position and its first n derivatives at time $t + \Delta t$. We then compute the forces (and thereby the accelerations) at the predicted positions. And then we find that these accelerations are *not* equal to the values that we had predicted. So we adjust our predictions for the accelerations to match the facts. But we do more than that. On the basis of the observed discrepancy between the predicted and observed accelerations, we also try to improve our estimate of the positions and the remaining $n-1$ derivatives. This is the "corrector" part of the predictor-corrector algorithm. The precise "recipe" used in applying this correction is a compromise between accuracy and stability. Here, we shall simply show a specific example of a predictor-corrector algorithm, without attempting to justify the form of the correctoı part.

Consider the Taylor expansion of the coordinate of a given particle aı time $t + \Delta t$:

$$r(t + \Delta t) = r(t) + \Delta t \frac{\partial r}{\partial t} + \frac{\Delta t^2}{2!}\frac{\partial^2 r}{\partial t^2} + \frac{\Delta t^3}{3!}\frac{\partial^3 r}{\partial t^3} + \cdots .$$