

The Random Walk and the Mean Squared Displacement

W. Smith and M. J. Gillan

Introduction

The origin of this article lies in a research project of ours, in which it became important to quantify the errors in the calculation of the mean squared displacement (MSD) - putting the error bars on a MSD plot. This set us thinking about the simple random walk, which offered a theoretical model for the calculations. The exercise proved to be interesting and enlightening, and despite the almost certain probability that this has been published elsewhere, we present our findings to readers of the CCP5 newsletter, who perhaps have not thought about these matters before.

We begin with the calculation of the MSD from a single time origin, and proceed to the case of multiple time origins. In both cases we derive formulae for the MSD and the associated errors. In the final section we put the theory to the test with a computer simulation.

Single time origin

Consider a 1 D random walk. A particle starts at location 0 (the origin), from where it makes a random 'hop' of distance d in the positive or negative direction, to a location $\pm d$. It can subsequently make any number of similar hops in succession and by so doing move quite some distance from the origin. It is important to note that the direction of a given hop is completely independent of any preceding hop, which is to say that the hops are completely *uncorrelated*. Effectively, this manner of motion, confines the particle to sites on a regular 1 D grid with spacing d .

We next ask the question: where will the particle be after n hops? This will be given by the formula

$$x_n = \sum_{i=1}^n h_i \quad (1)$$

which simply sums the contributions to the overall particle trajectory made by each hop h_i , where

$$h_i = \pm d. \quad (2)$$

So in order to determine the position x_n we need to know the history of the particle hops. We call such a history the *trajectory*, for obvious reasons. Normally a single trajectory is of little interest, as it is only one of a potentially astronomical number of possible trajectories that can result from a random walk of n hops. This being so, it is obvious that we need a statistical approach if we are to derive anything useful from such a model.

If we regard a single trajectory of n hops as one statistical ‘trial’, we can easily consider the outcome of an infinite number of such trials¹. We can represent this in the following way

$$\langle x_n \rangle = \left\langle \sum_{i=1}^n h_i \right\rangle, \quad (3)$$

where the angular brackets $\langle \dots \rangle$ indicate an average over an infinite number of trials. The result of this elegance is disappointing: the average $\langle x_n \rangle$ is zero, as can be seen from the following.

$$\begin{aligned} \langle x_n \rangle &= \left\langle \sum_{i=1}^n h_i \right\rangle \\ &= \sum_{i=1}^n \langle h_i \rangle \\ &= \sum_{i=1}^n 0 \\ &= 0. \end{aligned} \quad (4)$$

On reflection, this result is obvious: the i 'th step in any trajectory can be $+d$ or $-d$ with equal probability, so an average of an infinite number of i 'th hops must be zero. However the result is useful in that it shows how we can separate out contributions to the average, something we are allowed to do because each hop is independent of all others.

A more useful average is $\langle x_n^2 \rangle$, which of course is the *mean squared displacement*. We can calculate this as follows.

$$\begin{aligned} \langle x_n^2 \rangle &= \left\langle \left(\sum_{i=1}^n h_i \right)^2 \right\rangle \\ &= \left\langle \sum_{i=1}^n \sum_{j=1}^n h_i h_j \right\rangle \\ &= \left\langle \sum_{i=1}^n h_i^2 \right\rangle + \left\langle \sum_{i=1}^n \sum_{j \neq i}^n h_i h_j \right\rangle \\ &= \sum_{i=1}^n \langle h_i^2 \rangle + \sum_{i=1}^n \sum_{j=1}^n \langle h_i h_j \rangle \\ &= nd^2. \end{aligned} \quad (5)$$

¹Actually, this is a convenient fiction. While it is possible to imagine an infinite number of trials, there can only be 2^n possible independent trajectories. However, once n becomes sufficiently large, the conclusions drawn from the arguments above remain valid. We therefore assume that n is a large number throughout.

In obtaining this result we note that the loss of the cross terms ($i \neq j$) occurring after the fourth step above, results from the fact that the average $\langle h_i h_j \rangle$ must be zero - since the product $h_i h_j$ is $\pm d^2$, with equal probability for both signs, and once again an average of such terms must be zero.

The result (5) is of immediate interest to molecular simulators on at least two accounts. Firstly if it is imagined that each hop occurs after a regular time interval Δt , then it is apparent that n is directly proportional to the elapsed time t , which in turn means that the MSD is *linear in time*. Secondly in circumstances where the diffusional motion of atoms really is due to hops from one site to another, the model immediately reveals the average relationship between the hopping distance and the time interval between hops.

This simple model can be taken further. For example, a question often asked in molecular simulation is: how accurate is the MSD calculated in a simulation? The model offers insight into this issue also.

Since we know the MSD, as given by equation (5), we can proceed to determine the uncertainty in the result by calculating the *variance* for each point on the MSD, which we denote by σ_n^2 . This is given by the standard statistical formula:

$$\sigma_n^2 = \left\langle \left(x_n^2 - \langle x_n^2 \rangle \right)^2 \right\rangle, \quad (6)$$

which can be manipulated as follows

$$\begin{aligned} \sigma_n^2 &= \left\langle \left(x_n^2 \right)^2 \right\rangle - 2 \langle x_n^2 \rangle \langle x_n^2 \rangle + \langle x_n^2 \rangle^2 \\ &= \langle x_n^4 \rangle - \langle x_n^2 \rangle^2, \end{aligned} \quad (7)$$

where

$$\langle x_n^4 \rangle = \left\langle \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \sum_{\ell=1}^n h_i h_j h_k h_\ell \right\rangle \quad (8)$$

and (of course)

$$\langle x_n^2 \rangle^2 = (nd^2)^2. \quad (9)$$

Computing the term on the right of (8) is greatly simplified by recognising that since h_i , h_j , h_k , and h_ℓ , are uncorrelated when $i \neq j \neq k \neq \ell$, the only terms that survive the averaging process are those with:

- (a) $i = j = k = \ell$;
- (b) $(i = j) \neq (k = \ell)$;
- (c) $(i = k) \neq (j = \ell)$;
- (d) $(i = \ell) \neq (j = k)$;

where we note that the indices are equivalenced *in pairs* to guarantee a positive product for $h_i h_j h_k h_\ell$. We also note that the conditions (b), (c) and (d) are equivalent.

Thus from (a) we obtain the result:

$$\left\langle \sum_{i=1} h_i^4 \right\rangle = nd^4, \quad (10)$$

and from (b), (c) and (d) we obtain:

$$\left\langle \sum_{i=1} \sum_{j=1} h_i^2 h_j^2 \right\rangle = (nd^2)^2, \quad (11)$$

from which it follows that

$$\langle x_n^4 \rangle = (3n^2 + n)d^4, \quad (12)$$

or if n is extremely large, we may approximate this by

$$\langle x_n^4 \rangle = 3n^2 d^4. \quad (13)$$

Finally, combining this result with (9) gives

$$\sigma_n^2 = 2n^2 d^4. \quad (14)$$

We may thus quote the error in the MSD of the random walk as $\pm\sigma_n$, with

$$\sigma_n = \sqrt{2}nd^2. \quad (15)$$

This is an interesting result. It shows that the error in the MSD is directly proportional to the MSD itself (at least for a random walk). It is also quite large! However, this outcome is easily improved upon by calculating the MSD of many particles simultaneously. Then if the particles are independent, it is easy to show that the MSD remains the same, but the error is reduced to

$$\Sigma_n = \sqrt{\frac{2}{N}}nd^2, \quad (16)$$

where N is the number of particles. Even so 100 particles are necessary to improve the accuracy by an order of magnitude.

In 3 D, the corresponding results are easily shown to be

$$\begin{aligned} \langle r_n^2 \rangle &= 3nd^2, \\ \sigma_n &= \sqrt{6}nd^2, \\ \Sigma_n &= \sqrt{\frac{6}{N}}nd^2. \end{aligned} \quad (17)$$

While these results are interesting in themselves, they do not fully reflect how MSDs are calculated in simulations. The most obvious difference from practice, is that the results are obtained from a *single time origin* and consequently each point on a trajectory contributes only once to the overall averaging process. In practice it is usual to use many points on a trajectory, each as a time origin in its own right. The MSD is then calculated as an average over origins, as well as an average over trajectories. What are the statistical consequences of this?

Multiple time origins

For a single particle making M hops in 1 D, (where M is assumed to be a very large number,) we may take a point s on the trajectory as an origin and calculate the displacement of the particle from this origin after a further n hops using the formula

$$x_{n:s} = \sum_{i=s}^{n+s-1} h_i, \quad (18)$$

which is seen to be a generalisation of equation (1). It is important to note that $x_{n:s}$ does not represent an absolute location of the particle on the grid, but its displacement from the point labelled s . The squared displacement is given by

$$x_{n:s}^2 = \left(\sum_{i=s}^{n+s-1} h_i \right)^2. \quad (19)$$

The average of $x_{n:s}^2$ over N_o origins taken from the same trajectory is given by the equation

$$\bar{x}_n^2 = \frac{1}{N_o} \sum_{s=1}^{N_o} \left(\sum_{i=s}^{n+s-1} h_i \right)^2. \quad (20)$$

(Note that we have not assumed that every point on the trajectory has to be used as an origin.) We now recall that this is the result of averaging over just one trajectory. The average for an infinite number of trajectories is

$$\begin{aligned} \langle \bar{x}_n^2 \rangle &= \left\langle \frac{1}{N_o} \sum_{s=1}^{N_o} \left(\sum_{i=s}^{n+s-1} h_i \right)^2 \right\rangle \\ &= \frac{1}{N_o} \sum_{s=1}^{N_o} \left\langle \left(\sum_{i=s}^{n+s-1} h_i \right)^2 \right\rangle \\ &= \frac{1}{N_o} \sum_{s=1}^{N_o} n d^2 \\ &= n d^2, \end{aligned} \quad (21)$$

which is precisely the result obtained previously. It is perhaps gratifying to discover that the process of averaging over origins does not alter the expected result, but is it more accurate? To find out we must again calculate the variance, which is given by the same formula as before:

$$\bar{\sigma}_n^2 = \left\langle \left(\bar{x}_n^2 \right)^2 \right\rangle - \left\langle \bar{x}_n^2 \right\rangle^2, \quad (22)$$

where

$$\left\langle \bar{x}_n^2 \right\rangle^2 = n^2 d^4, \quad (23)$$

and

$$\begin{aligned} \left\langle \left(\bar{x}_n^2 \right)^2 \right\rangle &= \left\langle \left(\frac{1}{N_o} \sum_{s=1}^{N_o} \left(\sum_{i=s}^{n+s-1} h_i \right)^2 \right)^2 \right\rangle \\ &= \frac{1}{N_o^2} \left\langle \sum_{s=1}^{N_o} \sum_{t=1}^{N_o} \sum_{i=s}^{n+s-1} \sum_{j=s}^{n+s-1} \sum_{k=t}^{n+t-1} \sum_{\ell=t}^{n+t-1} h_i h_j h_k h_\ell \right\rangle. \end{aligned} \quad (24)$$

Once again we can reduce this expression if we recognise that h_i , h_j , h_k , and h_ℓ , are uncorrelated when $i \neq j \neq k \neq \ell$ and the only terms that survive the averaging process are those with:

- (a) $i = j = k = \ell$;
- (b) $(i = j) \neq (k = \ell)$;
- (c) $(i = k) \neq (j = \ell)$;
- (d) $(i = \ell) \neq (j = k)$;

where we note that (c) and (d) give rise to equivalent terms, but (a) and (b) are distinct.

Applying the condition (a) we obtain

$$\left\langle \left(\bar{x}_n^2 \right)^2 \right\rangle_{(a)} = \frac{1}{N_o^2} \left\langle \sum_{s=1}^{N_o} \sum_{t=1}^{N_o} \sum_{i=\max(s,t)}^{\min(n+s-1, n+t-1)} h_i^4 \right\rangle, \quad (25)$$

in which the peculiar limits of the third summation specify the points shared by the two sub-trajectories originating at s and t respectively. This can be made more explicit if we separate terms for which $s = t$ from those for which $s \neq t$, giving

$$\left\langle \left(\bar{x}_n^2 \right)^2 \right\rangle_{(a)} = \frac{1}{N_o^2} \left\langle \sum_{s=1}^{N_o} \sum_{i=s}^{n+s-1} h_i^4 \right\rangle + \frac{2}{N_o^2} \left\langle \sum_{s=1}^{N_o-1} \sum_{t>s}^{N_o} \sum_{i=t}^{n+s-1} h_i^4 \right\rangle. \quad (26)$$

The first term on the right deals with points on the *same* sub-trajectory only, while the second term deals with points *shared* between *different* sub-trajectories. It is now that we must recognise an

important fact: if we are to calculate the true statistical error, we cannot admit into the calculation data that are correlated, which means we must use only those sub-trajectories that have no points in common. (Put another way, the average over origins must be an average over statistically independent quantities, as is the average over trajectories.) It is obvious that the second term is comprised entirely of such points and should be ignored. Thus the result of this part of the calculation is

$$\begin{aligned}
\left\langle \left(\bar{x}_n^2 \right)^2 \right\rangle_{(a)} &= \frac{1}{N_o^2} \left\langle \sum_{s=1}^{N_o} \sum_{i=s}^{n+s-1} h_i^4 \right\rangle \\
&= \frac{1}{N_o^2} \sum_{s=1}^{N_o} \sum_{i=s}^{n+s-1} \langle h_i^4 \rangle \\
&= \frac{1}{N_o^2} \sum_{s=1}^{N_o} \sum_{i=s}^{n+s-1} d^4 \\
&= \frac{nd^4}{N_o}.
\end{aligned} \tag{27}$$

Proceeding with condition (b) we obtain

$$\begin{aligned}
\left\langle \left(\bar{x}_n^2 \right)^2 \right\rangle_{(b)} &= \frac{1}{N_o^2} \left\langle \sum_{s=1}^{N_o} \sum_{t=1}^{N_o} \sum_{i=s}^{n+s-1} \sum_{k=t}^{n+t-1} h_i^2 h_k^2 \right\rangle \\
&= \frac{1}{N_o^2} \left\langle \left(\sum_{s=1}^{N_o} \sum_{i=s}^{n+s-1} h_i^2 \right)^2 \right\rangle \\
&= \frac{1}{N_o^2} \left(N_o n d^2 \right)^2 \\
&= n^2 d^4.
\end{aligned} \tag{28}$$

In this case there is no difficulty with shared points, the sub-trajectories are independent.

Condition (c) gives

$$\begin{aligned}
\left\langle \left(\bar{x}_n^2 \right)^2 \right\rangle_{(c)} &= \frac{1}{N_o^2} \left\langle \sum_{s=1}^{N_o} \sum_{t=1}^{N_o} \sum_{i=\max(s,t)}^{\min(n+s-1, n+t-1)} \sum_{j=\max(s,t)}^{\min(n+s-1, n+t-1)} h_i^2 h_j^2 \right\rangle \\
&= \frac{1}{N_o^2} \left\langle \sum_{s=1}^{N_o} \sum_{t=1}^{N_o} \left(\sum_{i=\max(s,t)}^{\min(n+s-1, n+t-1)} h_i^2 \right)^2 \right\rangle \\
&= \frac{1}{N_o^2} \left\langle \sum_{s=1}^{N_o} \left(\sum_{i=s}^{n+s-1} h_i^2 \right)^2 \right\rangle + \frac{2}{N_o^2} \left\langle \sum_{s=1}^{N_o-1} \sum_{t>s}^{N_o} \left(\sum_{i=t}^{n+s-1} h_i^2 \right)^2 \right\rangle.
\end{aligned} \tag{29}$$

Once again we recognise that the second term right of the above equation is concerned with correlated origins, and so is discarded. Thus proceeding we obtain

$$\begin{aligned} \left\langle \left(\bar{x}_n^2 \right)^2 \right\rangle_{(c)} &= \frac{1}{N_o^2} \left\langle \sum_{s=1}^{N_o} \left(\sum_{i=s}^{n+s-1} h_i^2 \right)^2 \right\rangle \\ &= \frac{1}{N_o^2} \left(N_o \left(n d^2 \right)^2 \right) \\ &= \frac{n^2 d^4}{N_o}. \end{aligned} \quad (30)$$

And finally, for condition (d) we have

$$\left\langle \left(\bar{x}_n^2 \right)^2 \right\rangle_{(d)} = \frac{n^2 d^4}{N_o}, \quad (31)$$

which follows from its equivalence to (c).

Combining the results (27), (28), (30) and (31) into (24) and substituting with (23) into (22) gives the final result for $\bar{\sigma}_n^2$:

$$\bar{\sigma}_n^2 = \frac{2n^2 d^4}{N_o}. \quad (32)$$

From this we obtain the error as

$$\bar{\sigma}_n = \sqrt{\frac{2}{N_o}} n d^2. \quad (33)$$

So the error in the MSD calculated using a sum over *statistically independent* origins is just what might have been expected: the error for a single time origin, divided by $\sqrt{N_o}$. The error in an N particle average is then

$$\bar{\Sigma}_n = \sqrt{\frac{2}{N_o N}} n d^2. \quad (34)$$

In 3 D, the corresponding results are

$$\begin{aligned} \left\langle \bar{r}_n^2 \right\rangle &= 3n d^2, \\ \bar{\sigma}_n &= \sqrt{\frac{6}{N_o}} n d^2, \\ \bar{\Sigma}_n &= \sqrt{\frac{6}{N_o N}} n d^2. \end{aligned} \quad (35)$$

As before, it is noticeable that the error is directly proportional to the MSD, which provides the following useful mnemonic formula

$$\bar{\Sigma}_n = \sqrt{\frac{2}{3N_o N}} \left\langle \bar{r}_n^2 \right\rangle \quad (36)$$

as a general formula for estimating the error in the MSD.

In the following section we test these formulae against a simulated random walk, but beforehand we need to make some additional comments.

Firstly, for a full trajectory comprised of M data points, then the maximum number of statistically independent origins for sub-trajectories of length n is

$$N_o = \text{Int}(M/n). \quad (37)$$

Thus the value of N_o appearing in the formulae for the error will differ according to where we are on the MSD curve. This will cause the error to grow more rapidly with increasing n than equations (36) and (35) at first suggest.

Secondly, while the formula (36) may be used to estimate the error in the general case, its use hinges on the assumption that the underlying diffusional motion is a random walk (i.e. the motions of the atoms are independent and uncorrelated). One cannot therefore quote the results of this formula as the error if the MSD is non-linear. The best that can be said if the curvature of the MSD is outside the bounds defined by these errors is that the MSD is *inconsistent* with the random walk model. This may be a significant observation in some circumstances.

Thirdly, we have been careful in our calculation of the MSD and the associated errors to use uncorrelated origins, meaning that for sub-trajectories of length n , the origins are taken n points apart. What would be the consequences of ignoring this requirement and simply using every possible data point on the full trajectory of M points to calculate these? It turns out that the MSD is the same (as one might have expected). The error calculated in this way however, is quite different. The derivation is tedious, but it follows the lines outlined above, and we simply quote the (3 D) results.

$$\begin{aligned} \langle \bar{r}_n^2 \rangle &= 3nd^2, \\ \bar{\Sigma}_n^* &= \frac{d^2}{N_o} \sqrt{\frac{n^3(4N_o' - n)}{N}}. \end{aligned} \quad (38)$$

(Where we have used the asterisk to distinguish this formula from (35).) In this case as many as $M - n$ points on the trajectory can be used as origins for a sub-trajectory of length n . Thus the relationship between the number of origins (N_o'), M and n is

$$N_o' = M - n. \quad (39)$$

We should also note that in the derivation of (38) we have made the assumption that $N_o' \geq n$. The equality $N_o' = n$ shows that $n = M/2$ is the longest sub-trajectory for which the analysis holds. This is not an unreasonable condition to impose, and it greatly simplifies the final formula.

The formula (38) is very different from (36), but what difference it makes in practice is best shown by an example in the next section.

Example simulation

As an example we have performed a random walk simulation using a system of 100 ‘atoms’ with cubic periodic boundaries. The atoms had no mutual interaction and the trajectory of each is therefore completely independent. Each atom was initially placed in a random position in a $1 \times 1 \times 1$ cubic cell from which it underwent 16400 random hops in 3 D. Each hop was ± 0.1 units in the x, y and z directions simultaneously. The random number generator used for this purpose was that of Marsaglia *et al* [1].

The resulting MSD is presented Figure 1, with error bars calculated using formula (36) as an example. It is easy to see that the plot is linear and precisely the magnitude predicted by the formula (35).

In Figure 2 we present the estimated errors calculated by a number of different methods:

- (a) the variance obtained from direct calculation (see below);
- (b) the error calculated using formula (38);
- (c) the error calculated using formula (35);
- (d) the error calculated using the ‘blocking method’ [2].

In order to calculate the errors according to (a) and (d) the atom-averaged contributions to each point on the MSD arising from every possible time origin were stored - thus for example for the j -th point on the MSD, $16400 - j$ contributions were stored. The straightforward variance of all these contributions constituted the error (a). In the case (d) the blocking method [2] was used to calculate the error. This method attempts to eliminate correlation from the data and at worst determines the *lower bound* of the error. In Figure 2 we have plotted the maximum error determined by this method along with the associated uncertainties plotted as error bars. (Using the maximum error is by no means the only option, but it has the merit of simplicity. An analytical fit of the error as a function of the ‘blocking factor’ would undoubtedly yield a better estimate.)

The results in Figure 2 show clearly that method (a) offers a very poor estimate of the error, as might be expected given its complete disregard of correlation. Methods (b) and (c) give very similar results, which at first glance is surprising, but a closer inspection of formula (38) reveals why. If it is assumed that the number N'_o dominates the formula, we can write the approximation:

$$\bar{\Sigma}_n^* \simeq nd^2 \sqrt{\frac{4n}{NM}}, \quad (40)$$

which closely resembles the formula (35), if the definition (37) is taken into account. By this reckoning, the formulae (35) and (38) should differ only by a factor of $\sqrt{3/2}$, as is borne out by Figure 2. We note that the error according to (35) is larger because it has no correlation in the origins.

Lastly, the error calculated by the blocking method is seen to be a similar order of magnitude as the theoretical estimates. One hoped for a closer agreement than this, but it must be noted that

the estimate of the error is itself subject to considerable uncertainty [2]. These errors grow rapidly with the ‘blocking index’ as the error bars in Figure 2, for some representative points, reveal. Nevertheless the results here show that the error obtained in this way agrees with the theoretical estimate to within the calculated uncertainties. Thus it follows that the blocking method can be used to obtain the error in a general case (i.e. not just for random walks.)

Conclusions

We have shown that estimates of the error in the MSD are easily obtained from the random walk model. Simple formule have been provided. The key requirement in calculating the error is to eliminate, as far as possible, any correlation from the data. The blocking method is recommended as a model-independent method.

References

- [1] G. Marsaglia, A. Zaman and W. W. Tsang, *Stats, and Prob. Letters* **8** (1990) 35.
- [2] H. Flyvberg and H. G. Petersen, *J. Chem. Phys.* **91** (1989) 461.

Figure 1

Mean Squared Displacement

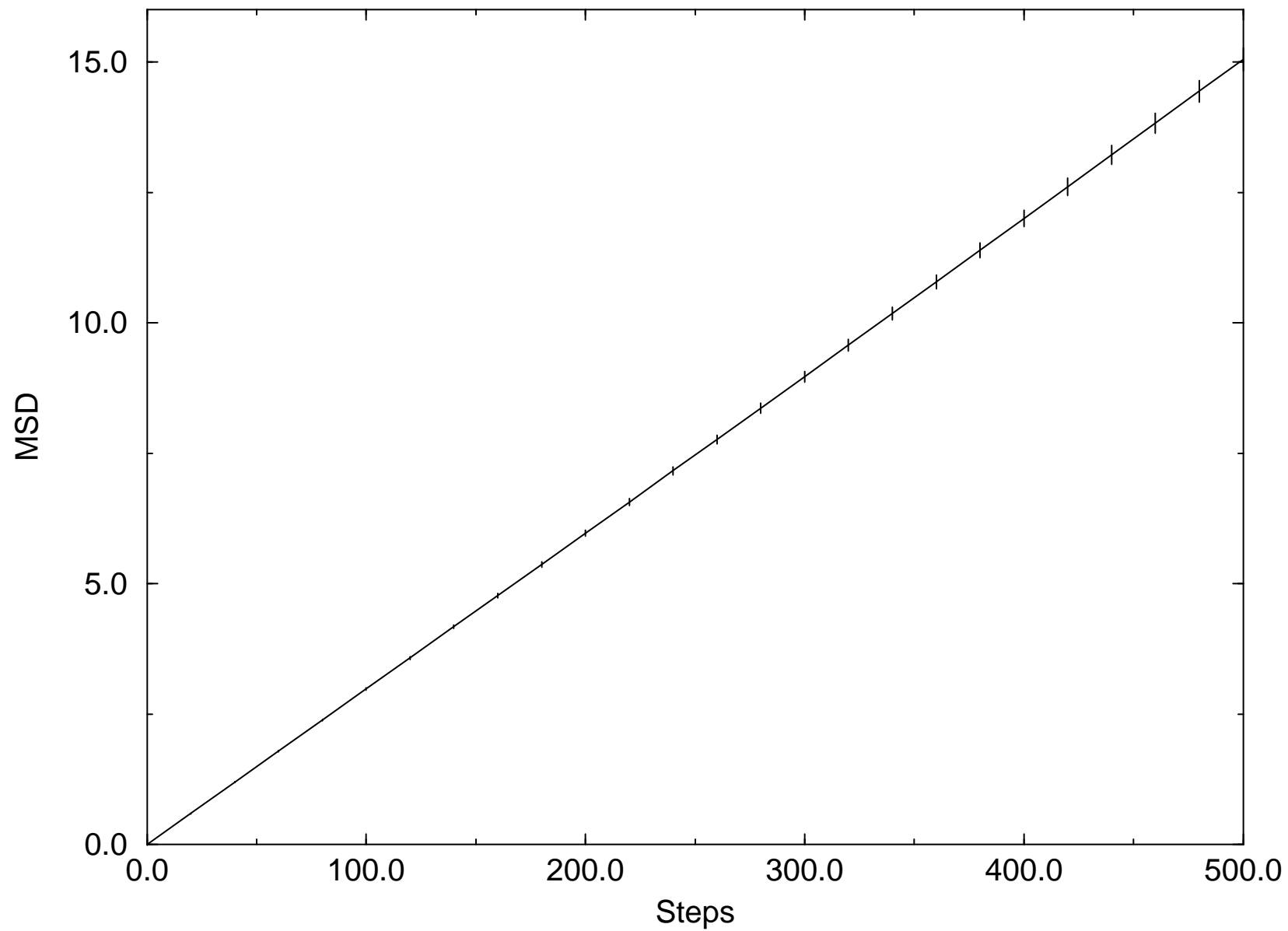
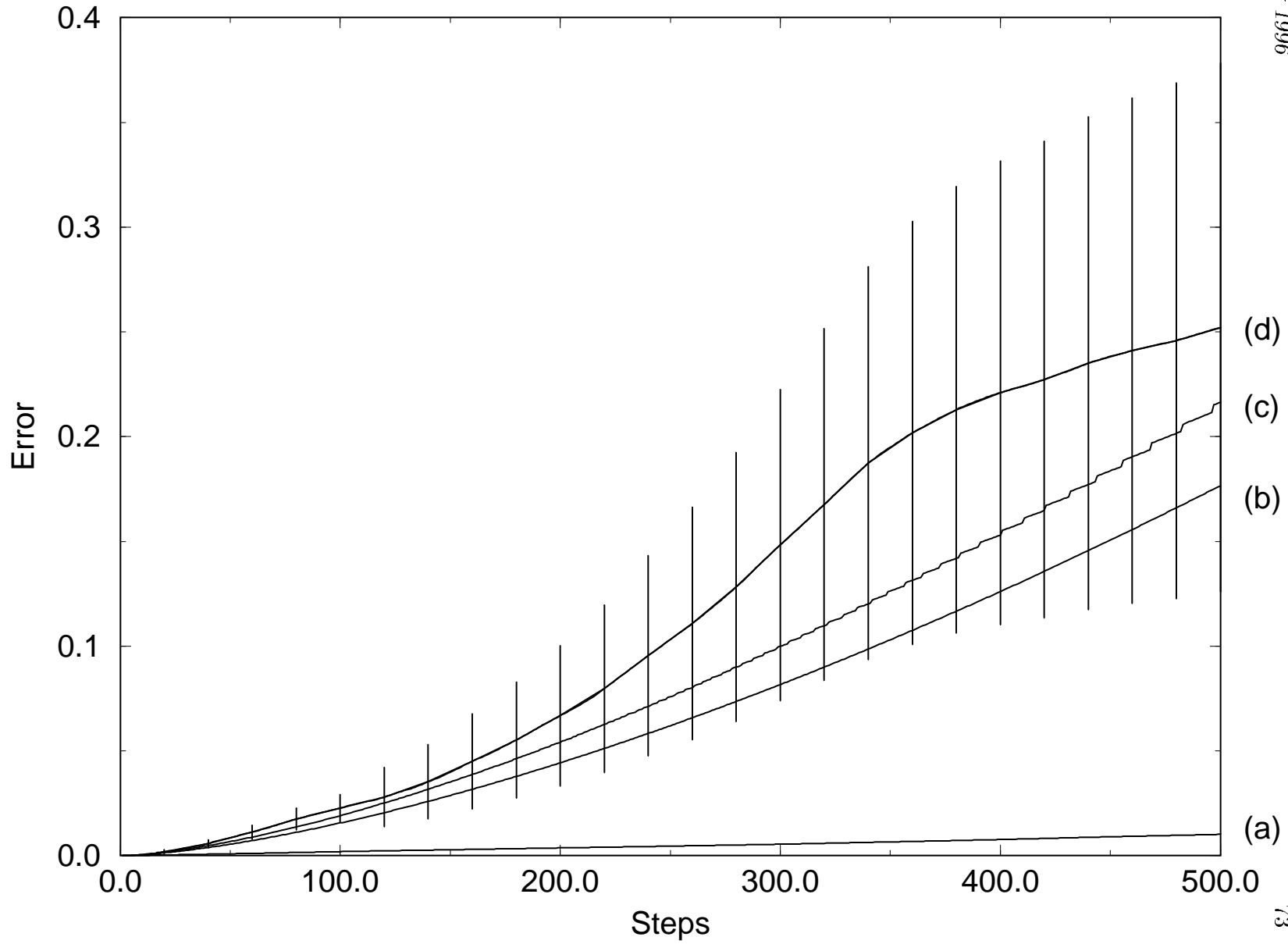


Figure 2

Estimated Errors



May 1996