Estimating Error in Diffusion Coefficients Derived from Molecular Dynamics Simulations

Gaurav Pranami and Monica H. Lamm*

Department of Chemical and Biological Engineering, Iowa State University, Ames, Iowa 50011, United States

Issues of Diffusion Coefficients Calculation from Molecular Simulations

- ❑ Statistical Uncertainty
- ❑ Observables with long-range or long-lived correlations
- ❑ Initial condition bias
- ❑ Finite simulation box size

Question based on those issues

- ❑ Length of the simulation to satisfy ergodic hypothesis
- ❑ Intervall between two successive samples
- ❑ Are Multiple Independent Simulations necessary

→ Not known a priori

Statistical Uncertainty

Sample mean $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$

Sample variance $S^2 = \frac{1}{N-1}\sum_{i=1}^{N} (X_i - \bar{X})^2$

Used to estimate population mean and
Variance from a sample of N independet
Measurements of the random variable X

The uncertainty associated with the estimate of population mean ($\bar{X}$) based on a one-sided Student's t-distribution is given by a $100(1-\alpha)$ confidence interval as follows:

$$\bar{X} \pm t_{N-1,1-\alpha/2}\frac{S}{\sqrt{N}}$$

Valid: if X is normally distributed or
if sample size is sufficiently large

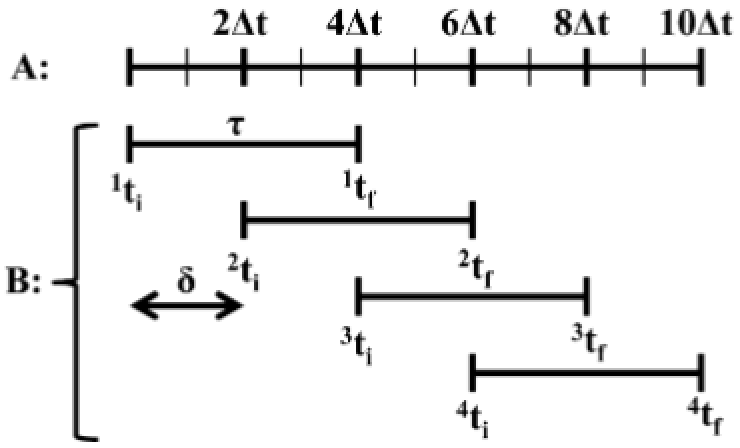
Errors with X are often reported as X +- S → Spread around estimated mean
95 % confidence Interval (alpha =0.05) → 95 % of intervals contain population mean

X± S /√N is often used (68,2 %)

Sample Correlation

Diffusion Coefficient. The diffusion coefficient (D) of a particle undergoing random walk (self-diffusion of LJ fluid and a rigid fractal aggregate diffusion in LJ fluid, as discussed later) is given by Einstein's relation[13]

$$D = \frac{1}{2d} \lim_{\tau \to \infty} \frac{d}{d\tau} \langle [\overrightarrow{r(\tau)} - \overrightarrow{r(0)}]^2 \rangle$$



$$\left\langle \left[\overrightarrow{r(\tau)} - \overrightarrow{r(0)}\right]^2 \right\rangle$$

$$\mathrm{MSD}(\tau, \delta) = \frac{1}{N_\tau} \sum_{j=0}^{N_\tau} \left(\frac{1}{N_\mathrm{P}} \sum_{i=1}^{N_\mathrm{P}} |\overrightarrow{r_i(j\delta + \tau)} - \overrightarrow{r_i(j\delta)}|^2 \right)$$

Know what you need.
Uncorrelated or Correlated sampling!

Sampling Squared Displacements

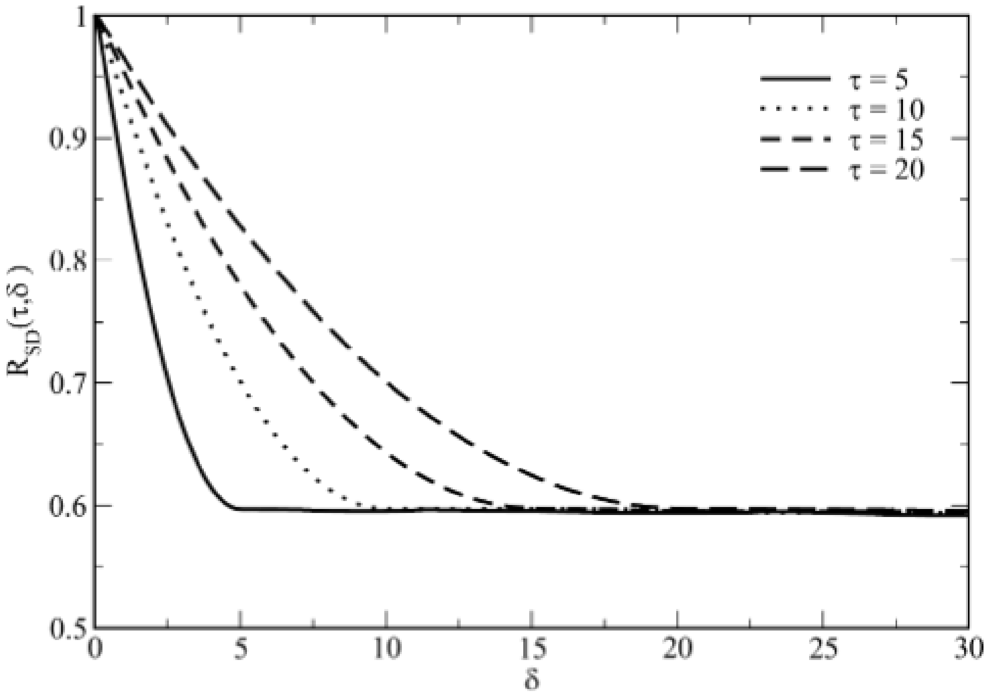
An accurate value of delta has to be determined to ensure Independent sampling

$$R_{\mathrm{SD}}(\tau,\,\delta) = \frac{\langle \mathrm{SD}(\tau,\,0)\;\mathrm{SD}(\tau,\,\delta)\rangle}{\langle \mathrm{SD}(\tau,\,0)\;\mathrm{SD}(\tau,\,0)\rangle}$$

$\Downarrow$

Decorrelates to a value of
$\delta \rightarrow \tau$

Should be sampled as non overlapping intervals



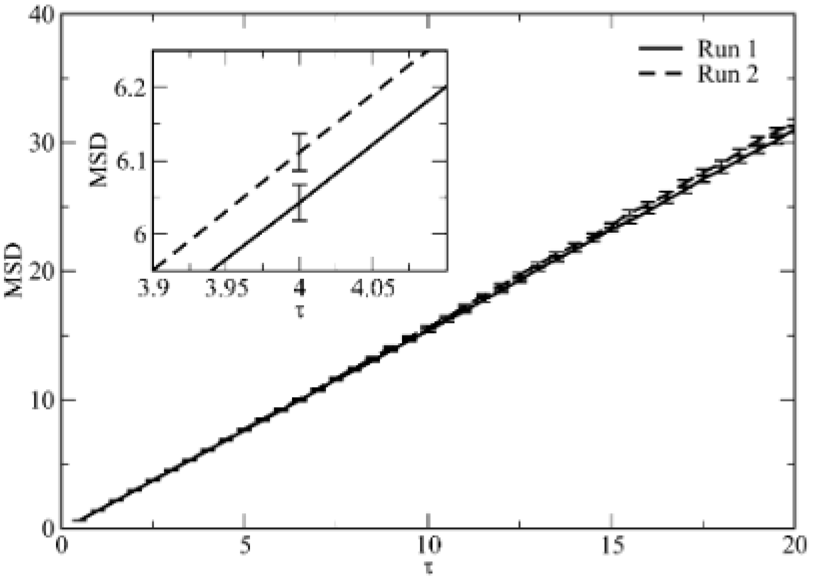
Linear Regression to calculate the Diffusion Coefficient

1/6 of the the slope of the linear fit of MSD
As a fucntion of tau

This would also produce the statistical
Uncertainty associated with the fitting
parameter

But this is only valid if

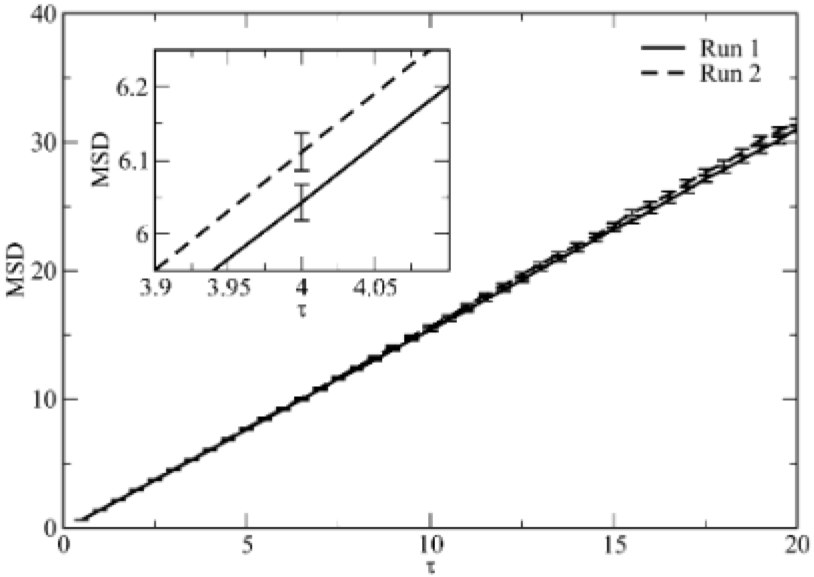
- ❑ Existence of SD having finite mean and variance
- ❑ Linear relationship between MSD and tau (Einstein‘s relation)
- ❑ Independence of SD samples
- ❑ Normal distribution of SD
- ❑ Constant variance in SD as a function of tau



Uncertainty in D can not be determined

Using Multiple Independent Simulations

Figure 4. Plot of MSD vs τ obtained from two independent MD simulations of an identical system of 125 LJ particles. The error bars indicate 95% confidence intervals. The mean squared displacements from these two runs are statistically different as indicated by nonoverlapping 95% confidence intervals. The inset shows a zoomed-in plot to highlight a representative nonoverlapping confidence interval, which is not visible in the main figure for lower values of τ.



An estimate of mean of D and its uncertainty using Students t-distribution can be obtained from multiple independent simulations

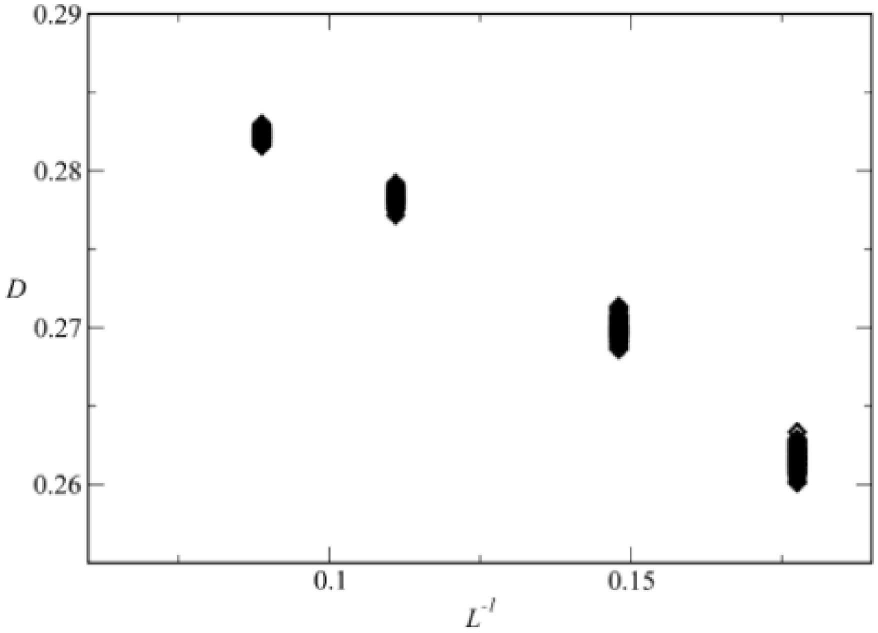
Or the confidence interval for the MSD can be determined invoking the central limit theorem if sample size ~>40.

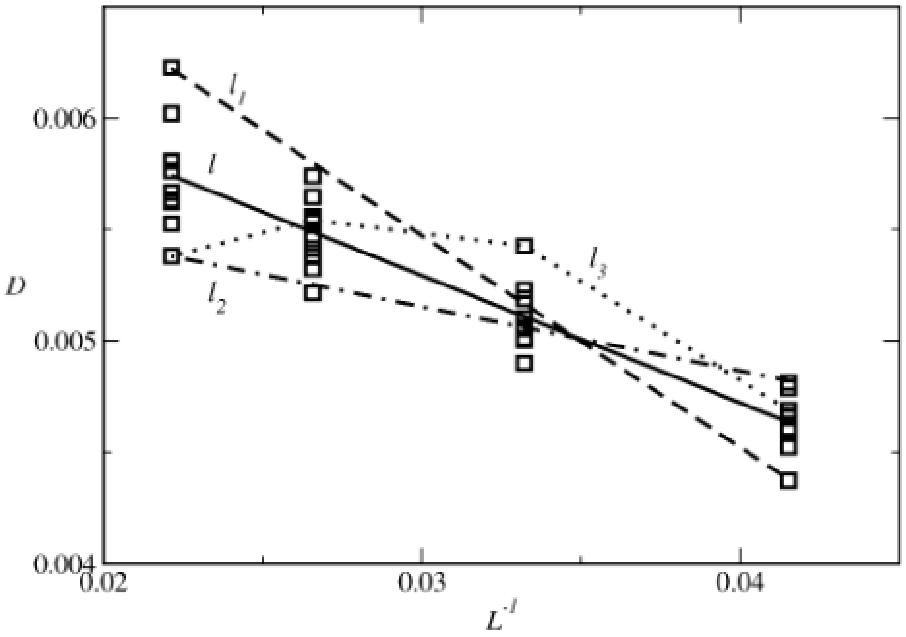
Diffusion as a Function of Simulation Box Length.

Motion of particles compromising a fluid induces a flow field leading to hydrodynamic interactions

Well known is this correctionfor the effects of a finite simulation box of size L

$$D_o = D + \frac{\xi k_b T}{6\pi\eta L}$$





Conclusion

Choose the correct sample interval and separation to achieve correlation or uncorrrelated sampling depending on requirement

Uncertainty of D can not be obtained from linear regression fitting as normal distribution and homoscedasticity are violated

But Muliple Independent Simulations can be used to calculate the uncertainty in D

Multiple Independent Calculations also show that the MSD from 2 MD simulations can be statistically different even for long simulations (simulation time not long enough to erase memory of intial state)