

Chem 860. Lecture 6

Algorithms for MD-III: Statistical error analysis

February 9, 2009

1 Analysis of statistical error

By statistical error, we mean that simulations may not be sufficiently long to contain enough statistically uncorrelated data to give accurate results. We imagine that results from different simulations distribute around the true value (for a given potential function), and our goal is to estimate the width of this distribution. We assume that the quantity that we are interested in follows Gaussian statistics, so that we only need to estimate the second moment. This Gaussian assumption is a reasonably good one, according to the central limit theorem, if the property is an average (sum) over many data points.

The most common quantity is expressed as the following average,

$$\langle A \rangle_{run} = A_\tau = \frac{1}{\tau} \int_0^\tau dt A(t)$$

One can then write the variance of A_τ as,

$$\sigma^2(A_\tau) = \langle A_\tau^2 \rangle - \langle A_\tau \rangle^2 = \frac{1}{\tau^2} \int_0^\tau \int_0^\tau dt dt' \langle [A(t) - \langle A \rangle][A(t') - \langle A \rangle] \rangle \quad (1)$$

One realize that the r.h.s. contains the auto-correlation function of A , thus

$$\sigma^2(A_\tau) = \frac{1}{\tau^2} \int_0^\tau \int_0^\tau dt dt' C_A(t - t') \quad (2)$$

Typically the autocorrelation function decays exponentially as a function of $t - t'$, i.e., $C_A(t - t') = C_A(0) \exp[-(t - t')/t^*]$, then we can further simplify Eq.2. Indeed, make a change of variables: $S = \frac{t+t'}{2}$; $s = t - t'$ (see for example, R. K. Pathria, p467, draw the integration square and integrate in a diagonal fashion),

$$\sigma^2(A_\tau) = \frac{1}{\tau^2} \left[\int_0^{\tau/2} dS \int_{-2S}^{2S} C_A(s) ds + \int_{\tau/2}^\tau dS \int_{2S-2\tau}^{2\tau-2S} C_A(s) ds \right] \quad (3)$$

In the limit that the observation time is much longer than t^* , we can further approximate the integral to be,

$$\sigma^2(A_\tau) = \frac{1}{\tau} \int_{-\infty}^{\infty} ds C_A(s) = \frac{2t^*}{\tau} C_A(0) = \frac{2t^*}{\tau} \sigma^2(A) \quad (4)$$

Generally, Eq.4 simply emphasizes that the variance in a measured quantity is inversely proportionally to the number of uncorrelated measurements (τ/t^*). If we **assume** that all the data points are uncorrelated, then $\tau/t^* \approx N$, thus we have

$$\sigma^2(A_\tau) = \frac{2}{N} \sigma^2(A) = \frac{2}{N} (\langle A^2 \rangle - \langle A \rangle^2)$$

The error is simply given by $\sigma(A_\tau)$. The result makes perfect sense: the fluctuation in the average is much smaller than the fluctuation in the data points! However, this simple estimate is too optimistic, because we **assumed** that all the data points are uncorrelated - which is not true - data points from MD simulations are likely to be correlated if they are not too far apart in time.

2 Block average

To better estimate errors, we use block average - which simply stated is a scheme that attempts to make sure that data points are uncorrelated. We separate data into n_b different blocks, with each block containing N_b data, so we have $N = n_b \times N_b$. We then compute the average over each block,

$$\langle A \rangle_b = \frac{1}{N_b} \sum_{i=1}^{N_b} A(t_i)$$

We then compute the second moments of $\langle A \rangle_b$, (note that the average is NOT affected by the blocking procedure)

$$\sigma^2(\langle A \rangle_b) = \frac{1}{n_b} \sum_{i=1}^{n_b} [(\langle A \rangle_b(i) - \langle A \rangle_{run})]^2$$

We repeat this procedure for different values of $N_b = 2, 4, 8 \dots$. If we plot $\sigma^2(\langle A \rangle_b)/n_b$ vs. N_b or n_b - which should reach a plateau when the blocked data points $\langle A \rangle_b$ become truly decorrelated from each other. In that case, we readily get

$$\sigma^2(\langle A \rangle_{run}) = \frac{1}{n_b} \sigma^2(\langle A \rangle_b)$$

When doing the block-average, you will appreciate how poor (or optimistic) the naive $1/N$ estimate is.

3 Size and length dependence of statistical error

- The error ($\sigma(\langle A \rangle)$) scales as $1/\sqrt{N}$ in terms of simulation length. So if one wants to decrease the error by a factor of 2, the length of simulation has to be extended by a factor of 4.
- For properties that can be written as a sum over particles in the system, i.e.,

$$\langle A \rangle = \left\langle \sum_{i=1}^M a_i \right\rangle = M \langle a \rangle$$

if one assumes that there is no correlation between different a_i , one can show straightforwardly that

$$\frac{\sigma^2 \langle A \rangle}{\langle A \rangle^2} \propto \frac{1}{M} \frac{\langle a^2 \rangle - \langle a \rangle^2}{\langle a \rangle^2}$$

So once again, the relative error ($\frac{\sigma \langle A \rangle}{\langle A \rangle}$) scales as $1/\sqrt{M}$. In other words, if one wants to decrease the error by a factor of 2, the size of simulation has to be extended by a factor of 4.