

Appendix C

Averages and fluctuations

C.1 Formulae for averaging

Note: this section was taken from ref [169].

When analyzing a MD trajectory averages $\langle x \rangle$ and fluctuations

$$\langle (\Delta x)^2 \rangle^{\frac{1}{2}} = \langle [x - \langle x \rangle]^2 \rangle^{\frac{1}{2}} \quad (\text{C.1})$$

of a quantity x are to be computed. The variance σ_x of a series of N_x values, $\{x_i\}$, can be computed from

$$\sigma_x = \sum_{i=1}^{N_x} x_i^2 - \frac{1}{N_x} \left(\sum_{i=1}^{N_x} x_i \right)^2 \quad (\text{C.2})$$

Unfortunately this formula is numerically not very accurate, especially when $\sigma_x^{\frac{1}{2}}$ is small compared to the values of x_i . The following (equivalent) expression is numerically more accurate

$$\sigma_x = \sum_{i=1}^{N_x} [x_i - \langle x \rangle]^2 \quad (\text{C.3})$$

with

$$\langle x \rangle = \frac{1}{N_x} \sum_{i=1}^{N_x} x_i \quad (\text{C.4})$$

Using eqns. C.2 and C.4 one has to go through the series of x_i values twice, once to determine $\langle x \rangle$ and again to compute σ_x , whereas eqn. C.1 requires only one sequential scan of the series $\{x_i\}$. However, one may cast eqn. C.2 in another form, containing partial sums, which allows for a sequential update algorithm. Define the partial sum

$$X_{n,m} = \sum_{i=n}^m x_i \quad (\text{C.5})$$

and the partial variance

$$\sigma_{n,m} = \sum_{i=n}^m \left[x_i - \frac{X_{n,m}}{m-n+1} \right]^2 \quad (\text{C.6})$$

It can be shown that

$$X_{n,m+k} = X_{n,m} + X_{m+1,m+k} \quad (\text{C.7})$$

and

$$\sigma_{n,m+k} = \sigma_{n,m} + \sigma_{m+1,m+k} + \left[\frac{X_{n,m}}{m-n+1} - \frac{X_{n,m+k}}{m+k-n+1} \right]^2 * \frac{(m-n+1)(m+k-n+1)}{k} \quad (\text{C.8})$$

For $n = 1$ one finds

$$\sigma_{1,m+k} = \sigma_{1,m} + \sigma_{m+1,m+k} + \left[\frac{X_{1,m}}{m} - \frac{X_{1,m+k}}{m+k} \right]^2 \frac{m(m+k)}{k} \quad (\text{C.9})$$

and for $n = 1$ and $k = 1$ (eqn. C.8) becomes

$$\sigma_{1,m+1} = \sigma_{1,m} + \left[\frac{X_{1,m}}{m} - \frac{X_{1,m+1}}{m+1} \right]^2 m(m+1) \quad (\text{C.10})$$

$$= \sigma_{1,m} + \frac{[X_{1,m} - mx_{m+1}]^2}{m(m+1)} \quad (\text{C.11})$$

where we have used the relation

$$X_{1,m+1} = X_{1,m} + x_{m+1} \quad (\text{C.12})$$

Using formulae (eqn. C.11) and (eqn. C.12) the average

$$\langle x \rangle = \frac{X_{1,N_x}}{N_x} \quad (\text{C.13})$$

and the fluctuation

$$\langle (\Delta x)^2 \rangle^{\frac{1}{2}} = \left[\frac{\sigma_{1,N_x}}{N_x} \right]^{\frac{1}{2}} \quad (\text{C.14})$$

can be obtained by one sweep through the data.

C.2 Implementation

In GROMACS the instantaneous energies $E(m)$ are stored in the energy file, along with the values of $\sigma_{1,m}$ and $X_{1,m}$. Although the steps are counted from 0, for the energy and fluctuations steps are counted from 1. This means that the equations presented here are the ones that are implemented. We give somewhat lengthy derivations in this section to simplify checking of code and equations later on.

C.2.1 Part of a Simulation

It is not uncommon to perform a simulation where the first part, *e.g.* 100 ps, is taken as equilibration. However, the averages and fluctuations as printed in the log file are computed over the whole simulation. The equilibration time, which is now part of the simulation, may in such a case invalidate the averages and fluctuations, because these numbers are now dominated by the initial drift towards equilibrium.

Using eqns. C.7 and C.8 the average and standard deviation over part of the trajectory can be computed as:

$$X_{m+1,m+k} = X_{1,m+k} - X_{1,m} \quad (\text{C.15})$$

$$\sigma_{m+1,m+k} = \sigma_{1,m+k} - \sigma_{1,m} - \left[\frac{X_{1,m}}{m} - \frac{X_{1,m+k}}{m+k} \right]^2 \frac{m(m+k)}{k} \quad (\text{C.16})$$

or, more generally (with $p \geq 1$ and $q \geq p$):

$$X_{p,q} = X_{1,q} - X_{1,p-1} \quad (\text{C.17})$$

$$\sigma_{p,q} = \sigma_{1,q} - \sigma_{1,p-1} - \left[\frac{X_{1,p-1}}{p-1} - \frac{X_{1,q}}{q} \right]^2 \frac{(p-1)q}{q-p+1} \quad (\text{C.18})$$

Note that implementation of this is not entirely trivial, since energies are not stored every time step of the simulation. We therefore have to construct $X_{1,p-1}$ and $\sigma_{1,p-1}$ from the information at time p using eqns. C.11 and C.12:

$$X_{1,p-1} = X_{1,p} - x_p \quad (\text{C.19})$$

$$\sigma_{1,p-1} = \sigma_{1,p} - \frac{[X_{1,p-1} - (p-1)x_p]^2}{(p-1)p} \quad (\text{C.20})$$

C.2.2 Combining two simulations

Another frequently occurring problem is, that the fluctuations of two simulations must be combined. Consider the following example: we have two simulations (A) of n and (B) of m steps, in which the second simulation is a continuation of the first. However, the second simulation starts numbering from 1 instead of from $n+1$. For the partial sum this is no problem, we have to add $X_{1,n}^A$ from run A:

$$X_{1,n+m}^{AB} = X_{1,n}^A + X_{1,m}^B \quad (\text{C.21})$$

When we want to compute the partial variance from the two components we have to make a correction $\Delta\sigma$:

$$\sigma_{1,n+m}^{AB} = \sigma_{1,n}^A + \sigma_{1,m}^B + \Delta\sigma \quad (\text{C.22})$$

if we define x_i^{AB} as the combined and renumbered set of data points we can write:

$$\sigma_{1,n+m}^{AB} = \sum_{i=1}^{n+m} \left[x_i^{AB} - \frac{X_{1,n+m}^{AB}}{n+m} \right]^2 \quad (\text{C.23})$$

and thus

$$\sum_{i=1}^{n+m} \left[x_i^{AB} - \frac{X_{1,n+m}^{AB}}{n+m} \right]^2 = \sum_{i=1}^n \left[x_i^A - \frac{X_{1,n}^A}{n} \right]^2 + \sum_{i=1}^m \left[x_i^B - \frac{X_{1,m}^B}{m} \right]^2 + \Delta\sigma \quad (\text{C.24})$$

or

$$\begin{aligned} & \sum_{i=1}^{n+m} \left[(x_i^{AB})^2 - 2x_i^{AB} \frac{X_{1,n+m}^{AB}}{n+m} + \left(\frac{X_{1,n+m}^{AB}}{n+m} \right)^2 \right] - \\ & \sum_{i=1}^n \left[(x_i^A)^2 - 2x_i^A \frac{X_{1,n}^A}{n} + \left(\frac{X_{1,n}^A}{n} \right)^2 \right] - \\ & \sum_{i=1}^m \left[(x_i^B)^2 - 2x_i^B \frac{X_{1,m}^B}{m} + \left(\frac{X_{1,m}^B}{m} \right)^2 \right] = \Delta\sigma \end{aligned} \quad (\text{C.25})$$

all the x_i^2 terms drop out, and the terms independent of the summation counter i can be simplified:

$$\begin{aligned} & \frac{(X_{1,n+m}^{AB})^2}{n+m} - \frac{(X_{1,n}^A)^2}{n} - \frac{(X_{1,m}^B)^2}{m} - \\ & 2 \frac{X_{1,n+m}^{AB}}{n+m} \sum_{i=1}^{n+m} x_i^{AB} + 2 \frac{X_{1,n}^A}{n} \sum_{i=1}^n x_i^A + 2 \frac{X_{1,m}^B}{m} \sum_{i=1}^m x_i^B = \Delta\sigma \end{aligned} \quad (\text{C.26})$$

we recognize the three partial sums on the second line and use eqn. C.21 to obtain:

$$\Delta\sigma = \frac{(mX_{1,n}^A - nX_{1,m}^B)^2}{nm(n+m)} \quad (\text{C.27})$$

if we check this by inserting $m = 1$ we get back eqn. C.11

C.2.3 Summing energy terms

The `g_energy` program can also sum energy terms into one, *e.g.* potential + kinetic = total. For the partial averages this is again easy if we have S energy components s :

$$X_{m,n}^S = \sum_{i=m}^n \sum_{s=1}^S x_i^s = \sum_{s=1}^S \sum_{i=m}^n x_i^s = \sum_{s=1}^S X_{m,n}^s \quad (\text{C.28})$$

For the fluctuations it is less trivial again, considering for example that the fluctuation in potential and kinetic energy should cancel. Nevertheless we can try the same approach as before by writing:

$$\sigma_{m,n}^S = \sum_{s=1}^S \sigma_{m,n}^s + \Delta\sigma \quad (\text{C.29})$$

if we fill in eqn. C.6:

$$\sum_{i=m}^n \left[\left(\sum_{s=1}^S x_i^s \right) - \frac{X_{m,n}^S}{m-n+1} \right]^2 = \sum_{s=1}^S \sum_{i=m}^n \left[(x_i^s) - \frac{X_{m,n}^s}{m-n+1} \right]^2 + \Delta\sigma \quad (\text{C.30})$$

which we can expand to:

$$\begin{aligned} & \sum_{i=m}^n \left[\sum_{s=1}^S (x_i^s)^2 + \left(\frac{X_{m,n}^S}{m-n+1} \right)^2 - 2 \left(\frac{X_{m,n}^S}{m-n+1} \sum_{s=1}^S x_i^s + \sum_{s=1}^S \sum_{s'=s+1}^S x_i^s x_i^{s'} \right) \right] \\ & - \sum_{s=1}^S \sum_{i=m}^n \left[(x_i^s)^2 - 2 \frac{X_{m,n}^s}{m-n+1} x_i^s + \left(\frac{X_{m,n}^s}{m-n+1} \right)^2 \right] = \Delta\sigma \end{aligned} \quad (\text{C.31})$$

the terms with $(x_i^s)^2$ cancel, so that we can simplify to:

$$\begin{aligned} & \frac{(X_{m,n}^S)^2}{m-n+1} - 2 \frac{X_{m,n}^S}{m-n+1} \sum_{i=m}^n \sum_{s=1}^S x_i^s - 2 \sum_{i=m}^n \sum_{s=1}^S \sum_{s'=s+1}^S x_i^s x_i^{s'} - \\ & \sum_{s=1}^S \sum_{i=m}^n \left[-2 \frac{X_{m,n}^s}{m-n+1} x_i^s + \left(\frac{X_{m,n}^s}{m-n+1} \right)^2 \right] = \Delta\sigma \end{aligned} \quad (\text{C.32})$$

or

$$- \frac{(X_{m,n}^S)^2}{m-n+1} - 2 \sum_{i=m}^n \sum_{s=1}^S \sum_{s'=s+1}^S x_i^s x_i^{s'} + \sum_{s=1}^S \frac{(X_{m,n}^s)^2}{m-n+1} = \Delta\sigma \quad (\text{C.33})$$

If we now expand the first term using eqn. C.28 we obtain:

$$- \frac{\left(\sum_{s=1}^S X_{m,n}^s \right)^2}{m-n+1} - 2 \sum_{i=m}^n \sum_{s=1}^S \sum_{s'=s+1}^S x_i^s x_i^{s'} + \sum_{s=1}^S \frac{(X_{m,n}^s)^2}{m-n+1} = \Delta\sigma \quad (\text{C.34})$$

which we can reformulate to:

$$- 2 \left[\sum_{s=1}^S \sum_{s'=s+1}^S X_{m,n}^s X_{m,n}^{s'} + \sum_{i=m}^n \sum_{s=1}^S \sum_{s'=s+1}^S x_i^s x_i^{s'} \right] = \Delta\sigma \quad (\text{C.35})$$

or

$$- 2 \left[\sum_{s=1}^S X_{m,n}^s \sum_{s'=s+1}^S X_{m,n}^{s'} + \sum_{s=1}^S \sum_{i=m}^n x_i^s \sum_{s'=s+1}^S x_i^{s'} \right] = \Delta\sigma \quad (\text{C.36})$$

which gives

$$- 2 \sum_{s=1}^S \left[X_{m,n}^s \sum_{s'=s+1}^S \sum_{i=m}^n x_i^{s'} + \sum_{i=m}^n x_i^s \sum_{s'=s+1}^S x_i^{s'} \right] = \Delta\sigma \quad (\text{C.37})$$

Since we need all data points i to evaluate this, in general this is not possible. We can then make an estimate of $\sigma_{m,n}^S$ using only the data points that are available using the left hand side of eqn. C.30. While the average can be computed using all time steps in the simulation, the accuracy of the fluctuations is thus limited by the frequency with which energies are saved. Since this can be easily done with a program such as xmgr this is not built-in in GROMACS.