Day 4 - Calculation of statistical errors

July 2024

1 Block analysis

Block analysis (Flyvbjerg, Petersen; JCP 1989) is a method used to correctly estimate the error on the mean in the case of correlated time series.

Recompute the average potential energies (column 3) as done in day 2 (for different temperatures from 0 to 3), together with the corresponding error on the mean estimated by the standard deviation (divided by square root of n. of samples) $\sigma_{\overline{x}}$. Remember to discard initial steps due to equilibration.

Then, write an algorithm to perform block analysis and use it to correctly estimate the error on the mean. Does this estimate agree with the one determined at previous point?

Hint: focus firstly on the energies at a given temperature and perform block analysis by varying the block size (in logarithmic scale); plot the error as a function of the block size. What happens starting from block size = 1 and increasing it? Search for a plateau region, which identifies the optimal block size. Then, use this block size for all the temperatures (assumption); plot both $\sigma_{\overline{x}}$ and the error estimated with block analysis as a function of the temperature.

We assumed that the optimal block size is approximatively the same for all the temperatures. Is this assumption correct? For which temperatures it is not? Why?

Repeat block analysis for random white noise. What is the result of block analysis? Does it agree with the error on the mean estimated by the standard deviation $\sigma_{\overline{x}}$?

2 Bootstrap on blocks

Bootstrap (Efron, 1979) is a procedure to compute properties of an estimator by random re-sampling with replacement from the data. Write an algorithm to perform boostrap on blocks:

1. split your time series into N (consecutive and disjoint) blocks, as done in block analysis (use the optimal N you have got in the previous point);

2. generate a dummy trajectory by randomly sampling N blocks with replacement (so that the trajectory has the same length as the original one) and use it to compute your quantity of interest (such as the mean energy or the heat capacity);

3. repeat 2nd step several times, then take average and standard deviation of the computed values; the standard deviation will estimate the uncertainty on the average value.

Use your algorithm to compute the error on the mean of the potential energy, as done in the previous point. Does this estimate agree with the previous one? Plot the two estimates as a function of the temperature.

Now, use your algorithm to compute the statistical error on the heat capacity, computed from fluctuations of the potential energy. Compare with the calculation of the heat capacity given by the derivative of the energy with respect to the temperature, approximated through finite difference.

What happens if you neglected to remove initial equilibration steps?

Finally, compare these values of heat capacity with those obtained by analysing the trajectories generated starting from the equilibrium configuration at $T = 3$ (hysteresis cycle). Where do you observe significative discrepancies? For such temperatures, did you correctly estimate the statistical error on the specific heat?