Lecture 8: Analysis of Simulations

Recap and Introduction

So far, we have covered the foundations of statistical mechanics and the two main sampling methods: Monte Carlo (MC) and Molecular Dynamics (MD). To run these simulations, we need to define the potential energy, $V(r)$, which can be done using either computationally intensive "ab initio" methods or more efficient force fields. We also use periodic boundary conditions (PBC) to simulate an infinite system.

A key question arises: **How do we know when to stop a simulation?** This leads to the topics of this lecture, which focus on the analysis of simulation data. We will cover:

- Stability and convergence
- Calculating averages and fluctuations
- Visualization of trajectories
- Correlation functions
- Transport properties

Simulation Stability

Errors and instabilities in a simulation can be caused by several factors:

- **Discretization:** The numerical integration of the equations of motion introduces errors.
- **Thermostats/Barostats:** If the coupling is too strong ("harsh") or too weak ("gentle"), it can lead to instability.
- **Large Timestep:** Using a timestep that is too large for the system's dynamics can cause the simulation to become unstable and "explode."
- **Phase Transitions:** The system may undergo a phase transition during the simulation.

A good way to check for stability is to monitor **conserved quantities**.

- In an **NVE** simulation, the total energy should be conserved.
- In **NVT** and **NPT** simulations, there are conserved quantities associated with the thermostat and barostat, respectively.
- Properties like temperature and pressure should also be conserved on average.

Averages and Fluctuations

The theoretical ensemble average of an observable, A, is given by $\langle A \rangle = \frac{1}{Z} \sum A_i e^{-\beta E_i}$. In practice, we compute this in two ways:

1. **Time Average:** We follow the trajectory of a single system and average the observable over time.

$$\langle A \rangle_t \approx \frac{1}{N} \sum_{i=1}^{N} A(\underline{x}(t_i), \underline{p}(t_i))$$

2. **Ensemble Average:** We imagine a vast collection ("ensemble") of identical systems and average the observable over all of them at a single point in time.

$$\langle A \rangle_\Gamma = \mathrm{Tr}(A\rho)$$

The **ergodicity hypothesis** states that for a system in equilibrium, these two averages are equivalent: $\langle A \rangle_t = \langle A \rangle_\Gamma$.

Fluctuations are a measure of the deviation from the average and are calculated as the variance:

$$\sigma_A^2 = \langle (A - \langle A \rangle)^2 \rangle = \langle A^2 \rangle - \langle A \rangle^2$$

These fluctuations can be related to measurable thermodynamic properties. For example, energy fluctuations in the NVT ensemble are related to the heat capacity at constant volume (C_V):

$$\sigma_E^2 = k_B T^2 C_V$$

It's important to note that the nature of fluctuations depends on the ensemble used, so they don't always have a direct, measurable physical meaning.

Equilibration and Convergence

When analyzing a simulation, we must first allow the system to reach **equilibrium**. Averages should only be calculated over the portion of the trajectory that comes after the system has equilibrated.

Because simulations have a finite length, there will always be some statistical uncertainty in the calculated averages. A potential pitfall is having the system get "stuck" in a small region of phase space, leading to poor sampling and inaccurate results. In complex cases, more advanced techniques like **block averaging** can be used to better estimate the uncertainty.

Visualization and Correlation Functions

Visualization is a crucial first step in analysis. Looking at the trajectory of atoms and molecules using software like VMD or ASE gives a qualitative understanding of the system's structure and dynamics.

A more quantitative analysis involves **correlation functions**, which measure how two properties are related to each other over time. The general form of a time correlation function is:

$$C_{AB}(t) = \langle A(t)B(0) \rangle$$

This function describes the average relationship between the value of observable A at a time t and the value of observable B at time $t = 0$. Because the system is in equilibrium, the correlation only depends on the time difference, not the absolute time.