Examination D-CHAB Fall 2023: Statistical Physics for CSE

19.08.2023 09:30-11:00

HIL G75

Prof. Sereina Riniker
Prof. Philippe H. Hünenberger

Problem statements in English

- Write your **name** and **identification (Legi) number** on **each** sheet of paper you hand in.
- The use of laptops, cell phones, calculators, books, course material, *etc.* is **not allowed** (exception: language dictionaries).
- You may use the **question sheets** to report your answers (or part of them).
- Please, clearly **highlight** (*e.g.* underline or frame) your **final answers** to the problems.
- Keep your answers **short**, but **clear**.
- The five problems of the examination have **equal weights** in the final mark.

1 Concepts and Understanding (F2023.1)

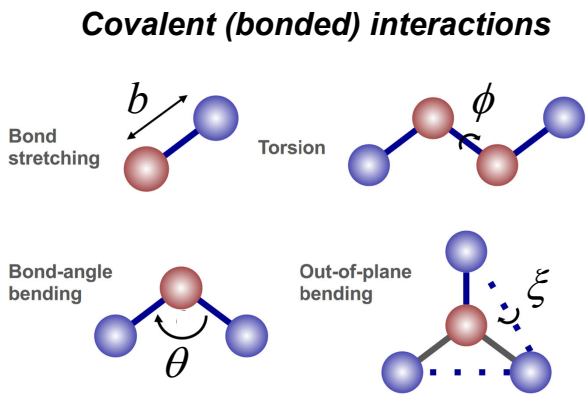
For each of the following items, **answer** the question(s) in a **clear** and **concise** way.

a. Give a precise statement of the central-limit theorem.

b. Consider a classical system with a coordinate vector $\mathbf{r}$ spanning a finite M-dimensional space Ω. The potential energy is $\mathcal{V}(\mathbf{r})$ and we consider an instantaneous observable $\mathcal{A}(\mathbf{r})$. We want to calculate numerically the canonical ensemble average $A = \langle \mathcal{A} \rangle_T$, where T is the absolute temperature. To this purpose, we sample the coordinate space using K points $\{\mathbf{r}_k \mid k = 1, .., K\}$ and monitor the corresponding values $\mathcal{V}_k = \mathcal{V}(\mathbf{r}_k)$ and $\mathcal{A}_k = \mathcal{A}(\mathbf{r}_k)$.
Write down the equations connecting these quantities to A in the limit $K \to \infty$, considering four alternative ways of performing the sampling:

- Systematic sampling, *i.e.* the K points are located on a uniform grid covering Ω.
- Stochastic sampling, *i.e.* we sample the K points at random following a uniform probability distribution over Ω.
- Isothermal molecular dynamics (MD), *i.e.* MD with a thermostat at temperature T.
- Umbrella sampling MD, *i.e.* isothermal MD using a biasing potential energy $\mathcal{V}_{bias}$, with corresponding values $\mathcal{V}_{bias,k} = \mathcal{V}_{bias}(\mathbf{r}_k)$ at the sampled points.

c. The drawing below depicts the four main types of covalent (bonded) interactions used in classical force fields. For each of these, write down a typical expression for the associated terms in the potential-energy function (as a function of the indicated quantities d, θ, ϕ and ξ).



2 Fundamental Equations (F2023.2)

For each of the following items, **write down** the relevant equation(s), **explain the meaning** of all the involved symbols, **state the SI units** of these quantities, and **answer** the additional questions.

a. Write down the Langevin equation of motion and answer the following questions:

- Explain in which context one uses the Langevin equation instead of the Newton one in molecular simulations.
- Give the name of the simulation method relying on the integration of the Langevin equation in time.
- Provide a physical interpretation for the different terms of the equation.
- Explain why the generated dynamics is not adiabatic (*i.e.* why energy flows in and out of the system).
- Explain why an isothermal situation is reached at equilibrium, and what balance determines the corresponding equilibrium temperature.
- Give the equation and the name of the simulation method corresponding to the limiting case of the Langevin equation for high friction.

b. Write the expression for the partition function Z of a molecular system in the canonical ensemble, both at the classical and at the quantum-mechanical level. Explain why Z is a central quantity in statistical mechanics. State the relationships (equations with a short explanation) connecting Z and its derivatives to the following important thermodynamic quantities in the canonical ensemble:

- The Helmholtz free energy.
- The internal energy.
- The entropy.

3 Derivations (F2023.3)

For each of the following items, **derive** the required expression analytically (*i.e.* it is not sufficient to only give the final result!), and **answer** the additional questions.

a. Consider a quantum-mechanical system with M energy states $m = 0, .., M-1$ and a statistical-mechanical ensemble encompassing K copies $k = 1, .., K$ of this system. The ensemble can be characterized by a population vector $\mathbf{p} = \{p_m \mid m = 0, .., M-1\}$ specifying the fraction of system copies that are in state m, with the normalization condition

$$\sum_{m=0}^{M-1} p_m = 1 \ .$$

The weight of a given population vector $\mathbf{p}$, *i.e.* the number of ways it can be achieved when distributing the K systems over the M states, is given by the multinomial coefficient

$$W(\mathbf{p}) = \begin{pmatrix} K \\ p_0, .., p_{M-1} \end{pmatrix} = \frac{K!}{\prod_{m=0}^{M-1}(Kp_m)!} \ .$$

The associated Boltzmann entropy is defined as

$$S(\mathbf{p}) = k_B K^{-1} \ln W(\mathbf{p}) \ ,$$

where k_B is the Boltzmann constant. Using these equations along with the Stirling approximation, show that in the limit of large K, S can be reformulated as a Gibbs entropy, *i.e.*

$$S(\mathbf{p}) = -k_B \sum_{m=0}^{M-1} p_m \ln p_m \ .$$

b. Consider two Hamiltonians $\mathcal{H}_A = \mathcal{H}_A(\mathbf{r}, \mathbf{p})$ and $\mathcal{H}_B = \mathcal{H}_A(\mathbf{r}, \mathbf{p})$ that depend on the Cartesian coordinate vector $\mathbf{r}$ and momentum vector $\mathbf{p}$ of a molecular system, and correspond to "alchemically" different states of the system (*e.g.* same molecule substituted with a chlorine atom in A and a bromine atom in B). In the context of the canonical ensemble, derive the free-energy perturbation (FEP) formula

$$\Delta F_{AB} = F_B - F_A = -k_B T \ln \left\langle e^{-\beta(\mathcal{H}_B - \mathcal{H}_A)} \right\rangle_A \ ,$$

where k_B is the Boltzmann constant and T the absolute temperature, which relates the Helmholtz free-energy difference ΔF_{AB} to an ensemble average $\langle .. \rangle_A$ calculated using the Hamiltonian $\mathcal{H}_A$. For this, you will need to perform the following steps: (*i*) write the equation connecting the Helmholtz free energy F to the canonical partition function Z; (*ii*) write the equation for the classical partition function Z (as determined by the Hamiltonian $\mathcal{H}$); (*iii*) use these two expressions to formulate ΔF_{AB}; (*iv*) recast the resulting expression in the form of a canonical ensemble average $\langle .. \rangle_A$.

4 Explicit Calculations (F2023.4)

For each of the following items, **calculate** the numerical result paying particular attention to the **units**, and **answer** the additional questions.

a. Consider the throw of $K = 4$ dice, whereby each dice has $M = 6$ faces. Assume first that the dice all have different colors, so that a throw is characterized by a sequence of numbers (*e.g.* $\{1, 3, 2, 6\}$ would stand for the faces shown by the red, green, blue and yellow dice), and answer the following questions:

   - Give the total number of possible distinguishable sequences for a throw.
   - Give the probabilities of the sequences $\{1, 1, 1, 1\}$, $\{1, 2, 2, 1\}$ and $\{1, 2, 4, 6\}$.
   - State what is (are) the most probable sequence(s).

   Now assume that the dice have all the same color, so that a throw is characterized instead by a distribution of face occurrences (*e.g.* $\{1, 0, 1, 0, 2, 0\}$ would indicate that one dice has value 1, one dice has value 3, and two dice have value 5), and answer the following questions:

   - Give the total number of possible distinguishable distributions for a throw.
   - Give the probabilities of the distributions $\{1, 1, 1, 1, 0, 0\}$, $\{2, 2, 0, 0, 0, 0\}$ and $\{4, 0, 0, 0, 0, 0\}$.
   - State what is (are) the most probable distribution(s).

b. A diatomic gas is stored in a container with a volume V of 1 liter at a temperature $\tilde{T}$ of 27 °C. We measure a pressure P of 74.7 bar. Assuming that the gas behaves ideally, calculate the number of atoms in the container. For this calculation, you can use the following approximate numerical values: $R = 8.3$ J·mol^{-1}·K^{-1} for the gas constant, $N_A = 6 \cdot 10^{23}$ mol^{-1} for the Avogadro number, and $\tilde{T}_o = -273$ °C for the absolute-zero temperature.

5 Algorithms and Implementation (F2023.5)

For each of the following items, **write the code** of a `C++` function performing the required task (or, at least, the pseudo-code; the exactness of your `C++` syntax will not be graded), and **answer** the additional questions.

a. Write a `C++` function `RmsdCalc` that will consider the coordinate trajectory of a macromolecule of N atoms in vacuum along a simulation of duration t_{sim}, and calculate the time series of the root-mean-square atomic positional deviation (RMSD) relative to a provided reference structure. The RMSD is calculated as

$$s(t) = \left\{ N^{-1} \sum_{n=1}^{N} [\tilde{\mathbf{r}}_n(t) - \mathbf{r}_n^*]^2 \right\}^{1/2} ,$$

where $\mathbf{r}_n^*$ is the Cartesian coordinate vector of atom n in the reference structure, and $\tilde{\mathbf{r}}_n(t)$ the corresponding vector $\mathbf{r}_n(t)$ in the trajectory frame at time t after roto-translational superimposition of this frame onto the reference structure ($\mathbf{r}_n(t) \rightarrow \tilde{\mathbf{r}}_n(t)$). In practice, the trajectory is discretized using a timestep Δt, so that time is specified by an integer index $k = 0, .., K-1$ as $t = k\Delta t$, where K is the the total number of trajectory frames (*i.e.* $K\Delta t = t_{sim}$). The function declaration reads

```
void RmsdCalc ( int N, int K, double ref[], double trj[][3*N], double rmsd[] );
```

Provided to the function are the number `N` of atoms, the number `K` of trajectory frames, the reference structure `ref` and the trajectory `trj`. The element `ref[3*n+i]` of `ref` contains the `i`th component (0,1,2 for x,y,z) of $\mathbf{r}^*$. Similarly, the element `trj[k][3*n+i]` of `trj` contains the `i`th component (0,1,2 for x,y,z) of $\mathbf{r}_n(k\Delta t)$. The function should fill the array `rmsd` with the values of $s(k\Delta t)$ for $k = 0, .., K-1$. It is assumed that a library function `RotoTransFit(ref,fit)` is available in your code for performing the roto-translational superimposition of a structure `fit` onto a structure `ref` (*i.e.* you can call the function and do not need to write the corresponding code).

Additional questions:

- Explain why the roto-translational superimpositon is necessary.
- When the simulation is performed in solution under periodic boundary conditions (rather than in vacuum), state what additional step has to be performed in order to calculate the RMSD, and whether it should be carried out before or after the roto-translational superimpositon.