Data exploration of molecular simulation trajectories

1 Data set

We have two trajectories of a particle (mass =1 kg) in a two-dimensional potential (traj1.txt and traj2.txt). The Trajectories contain position coordinates (in m) and the potential energy at that position (in J). One of the trajectories has been computed by a molecular dynamics simulation, the other by Metropolis–Hastings Monte-Carlo. Both trajectories have 1000000 time steps.

1.1 Visualization of trajectories

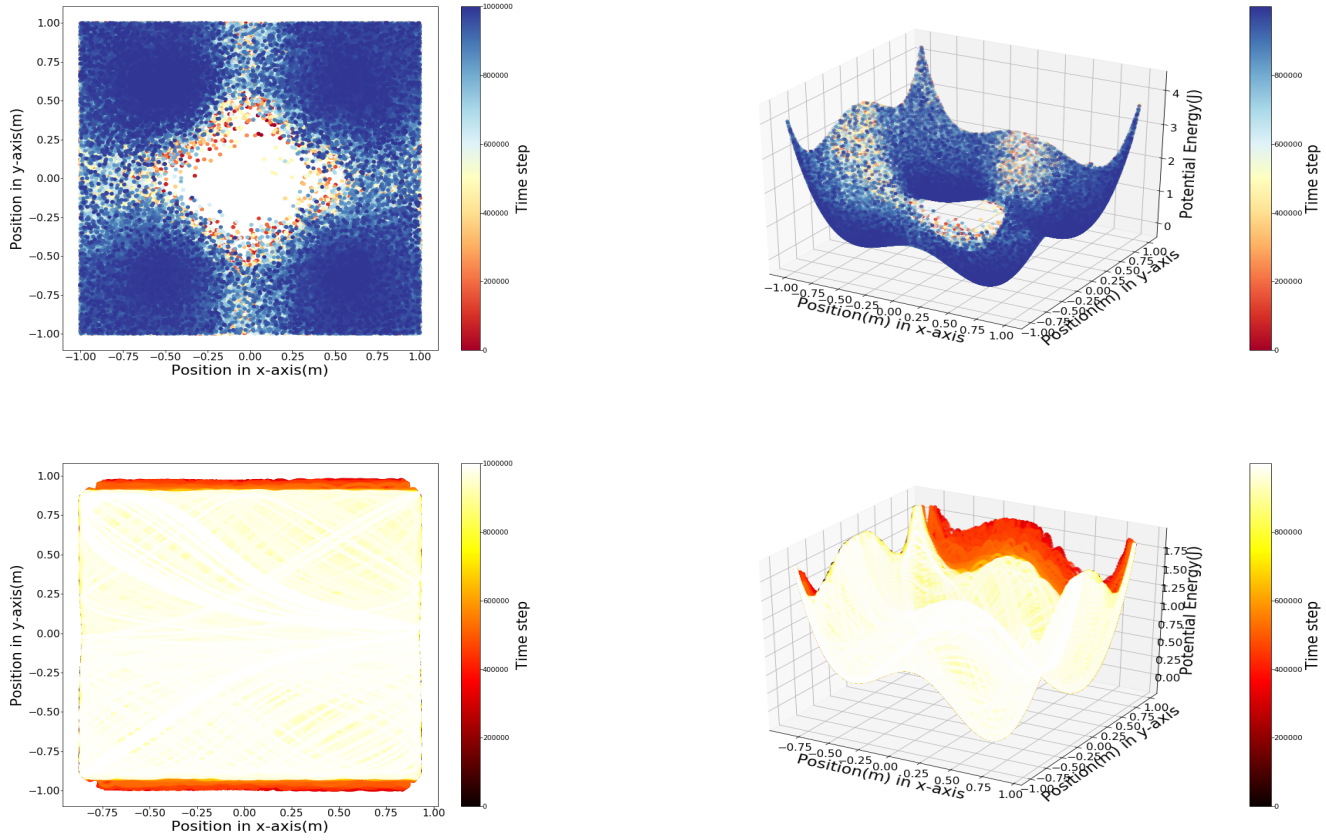

Figure 1.1: Trajectory 1 and 2

From figure 1.1, 1.2 and 1.3, we can draw following inferences.

- 1st Trajectory seems to be discontinuous and particle mostly stays in four regions in later stages of simulation.

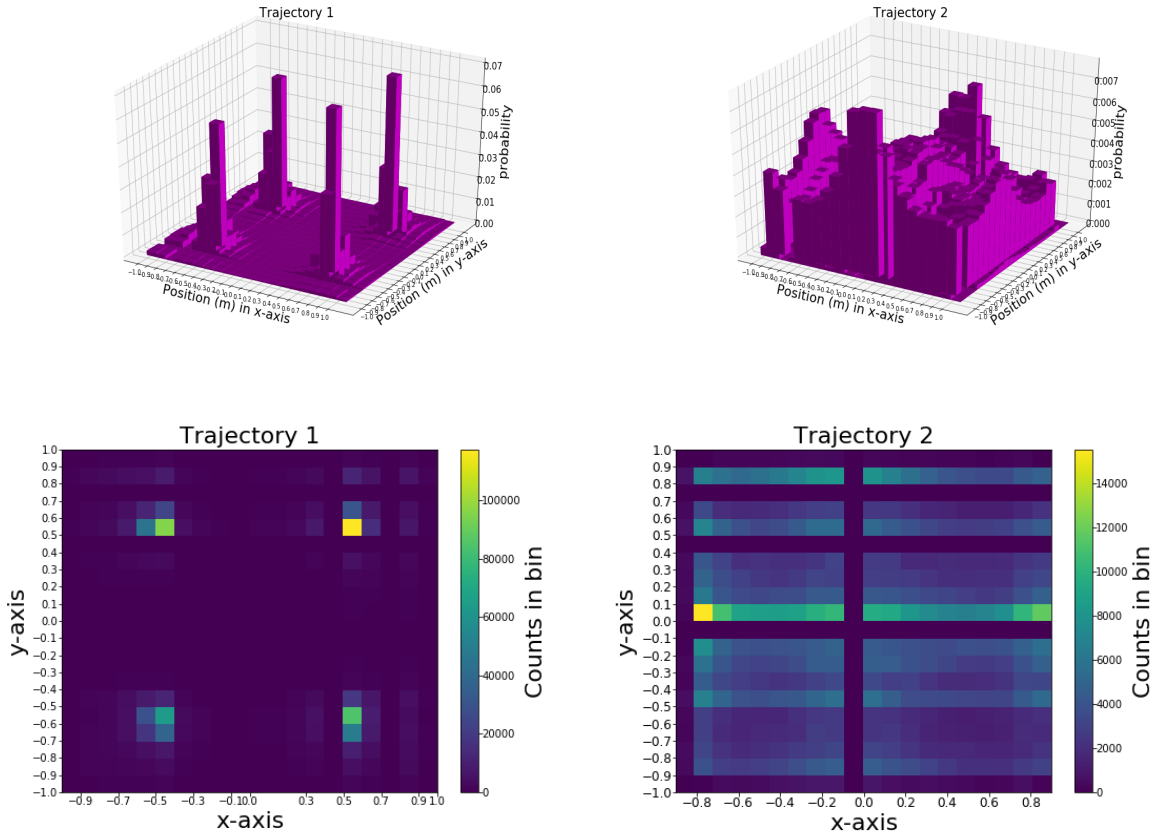

Figure 1.2: Histogram for Trajectory 1 and 2

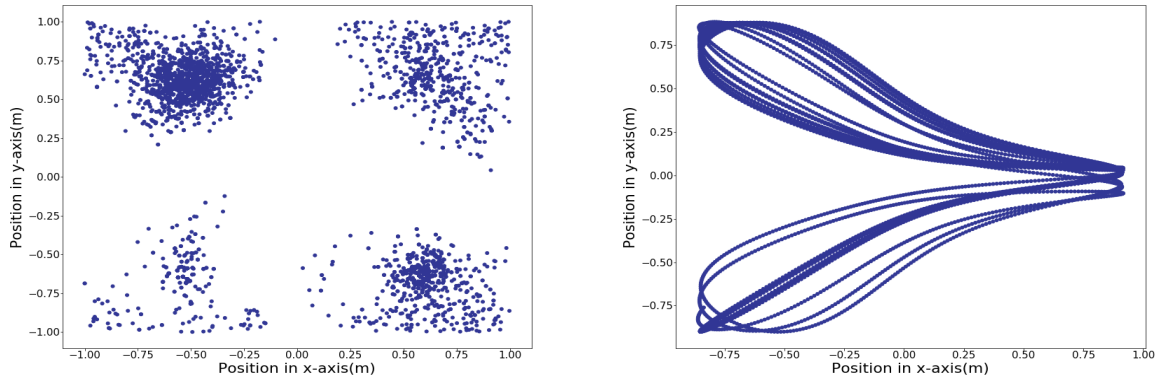

Figure 1.3: Plots of last 5000 points of trajectory 1 and 2

- The second is a continuous trajectory (from figure 1.3)and it is similar to a motion of a pendulum having around 0.8 m of displacement from equilibrium position.

1.2 Clustering of Trajectory

For clustering the positions, **K-means Clustering** algorithm is used. Before clustering, elbow method is also used to find out optimal number of clusters. The basic idea behind partitioning methods, such as k-means clustering, is to define clusters such that the total intra-cluster variation (or total within-cluster sum of square (WSS)) is minimized and we want it to be as small as possible. We compute clustering algorithm (e.g., k-means clustering) for different values of k (1 to 10). Figure 1.7 shows results of elbow method for both trajectories.

1.2.1 Elbow Method:

The Elbow Method can be defined as follow:

- For each k, calculate the total of within-cluster sum of square (wss).
- Plot the curve of wss according to the number of clusters k.
- The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

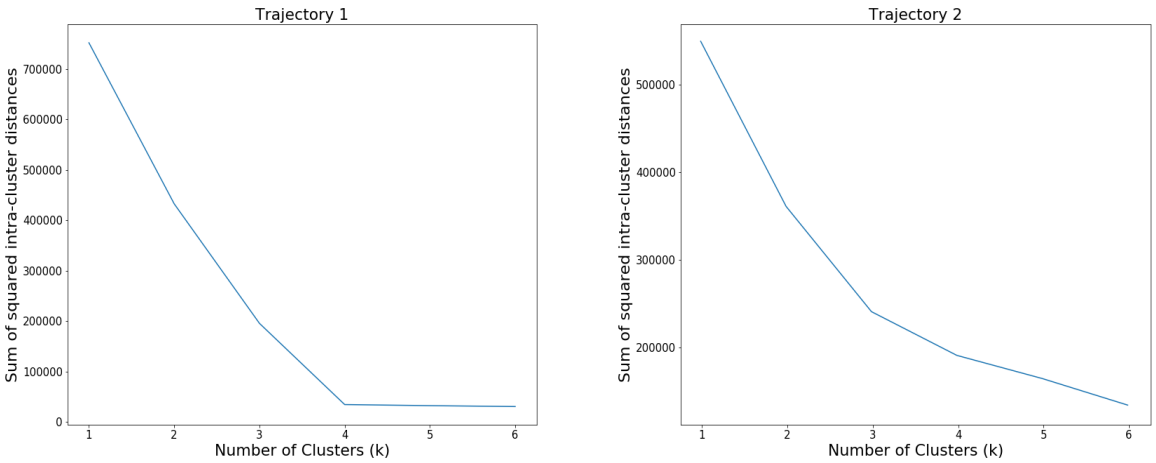

Figure 1.4: Elbow method for Trajectory 1 and 2

1.2.2 Inferences

For First trajectory, elbow method shows 4 clusters clearly, while for second trajectory it is not very much clear.

1.2.3 K-means Clustering :

- Randomly choose cluster centers.
- assign points to closest center.
- update centers
- reassign the points to cluster centers

- repeat until the convergence reached. The convergence means when the membership no longer changes or changes less than the tolerance limit

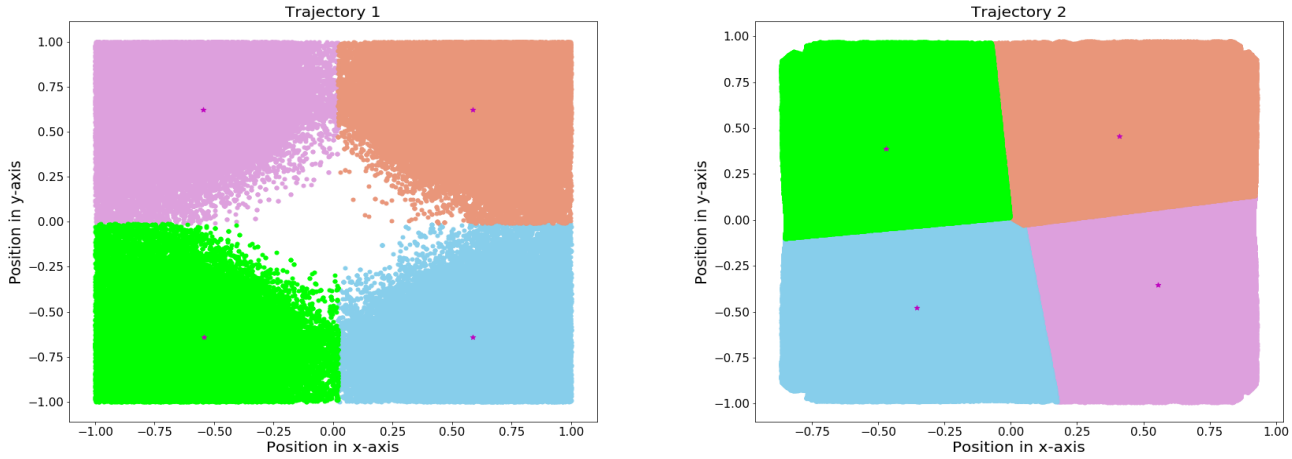

Figure 1.5: Clustering using K-means

Probability value of a cluster is ratio of number of points within the cluster and total no of trajectory points.
Probability values of clusters of trajectory 1 are [0.239406 0.255673 0.253509 0.251412]
Probability values of clusters of trajectory 2 are [0.22194 0.304687 0.224204 0.249169]

1.3 Analysis of method of generation for trajectories

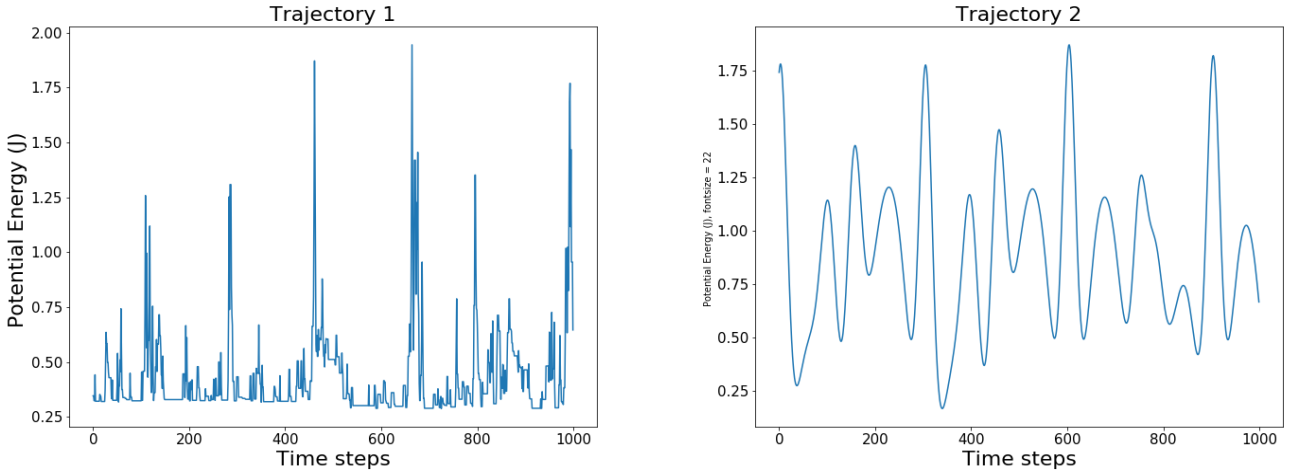

Figure 1.6: Potential for first 1000 positions

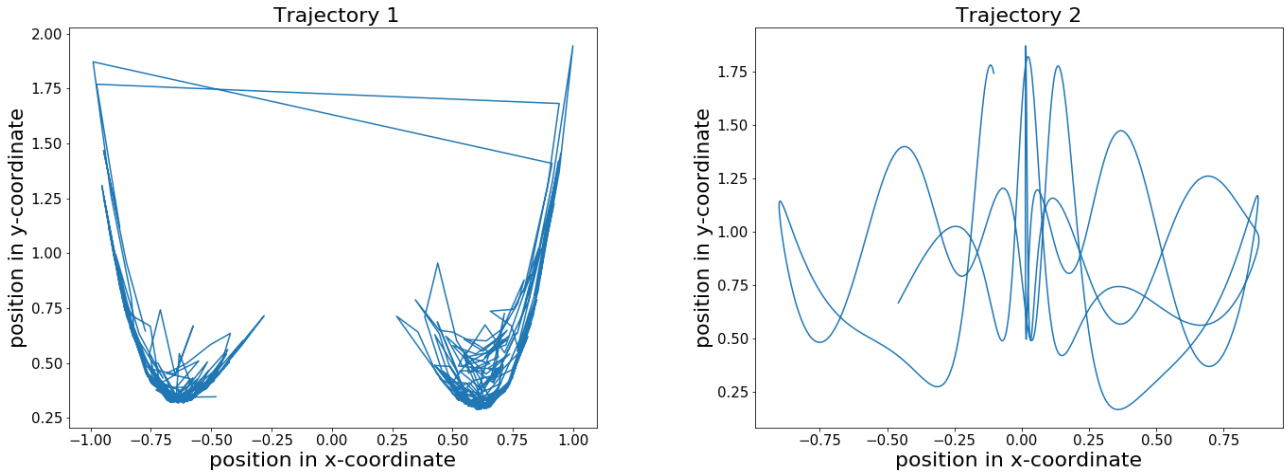

Figure 1.7: Position in XY plane for first 1000 points

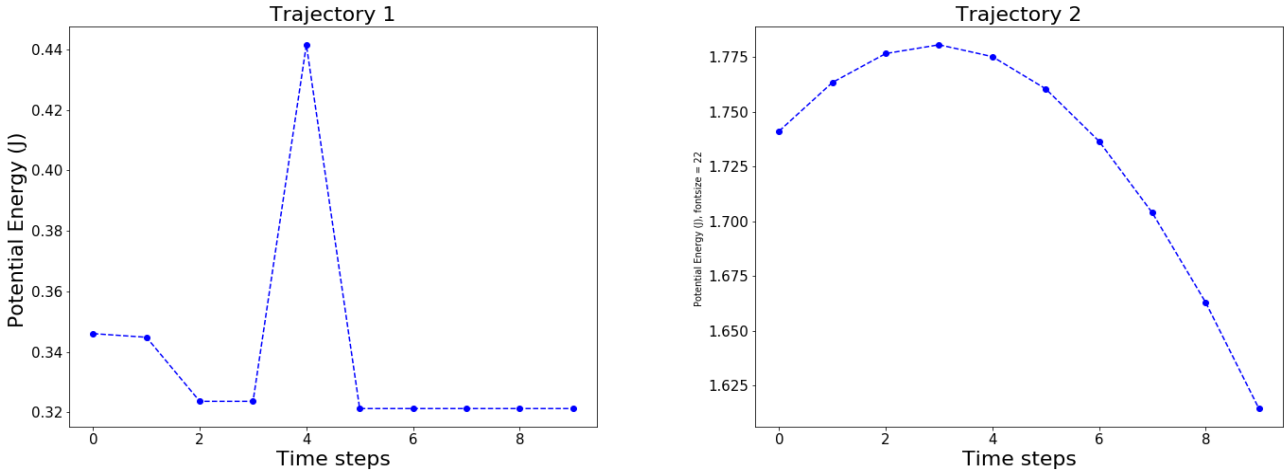

Figure 1.8: Potential for first 10 position

1.3.1 Inferences

- First Trajectory is generated by Metropolis–Hastings Monte-Carlo and second Trajectory is generated by Molecular Dynamics
- In first trajectory, many consecutive positions are same, this can only happen in case of Metropolis Hastings Monte carlo simulation, when the the proposal position is not accepted. while in Second trajectory, every consecutive entry is different and shows dynamics of particle. Thus, Trajectory 1 is generated by Metropolis–Hastings Monte Carlo while 2nd using MD simulations.
- From figure 1.9, for trajectory 2, potential is a continuous and differentiable function, so we can calculate forces at any point deterministic way. Trajectory 1 is continuous but not differentiable ,So we can not calculate force at a sharp points / or right sided derivative will be different from left sided.

- From Figure 1.10, for trajectory 2, position plot is smooth,so its continuous and differentiable , so we can calculate velocity at any point deterministically so we can go backwards if negative timescale is applied. While for trajectory 1, there are sharp edges and jumps in plot, so we forward velocity is different than backward, so we can not go backwards and the next state is dependent on the current state, so its metropolis Hastings Monte Carlo sampling.
- In Figure 1.11, the energy pattern for first 10 positions for both trajectories has been plotted. In trajectory 1, transitions 3-4 and 4-5 holds a high energy fluctuations in a single step which contradicts molecular dynamics of particle . While in 2nd trajectory, there are always small changes in energy in consecutive steps which shows particle undergoes small changes in velocities and other parameters. In MD, particle can undergo high energy changes(uphill or downhill) but in several steps only, However, it can happen in Monte carlo in a single step. Thus it shows 2nd trajectory is Molecular dynamics , while 1st is Metropolis Monte Carlo Hastings.

1.4 Visualisation of the potential and Relative free Energy

The potential are plotted by discretising the positions by 0.1 m grid size in each dimension using Scipy and cubic interpolation method. The relative free energy plots are calculated by $G - Gmax = -Kb * T \log(P/Pmax)$ where Kb*T=1 and is taken as zero for the regions where probability of finding particle is zero.

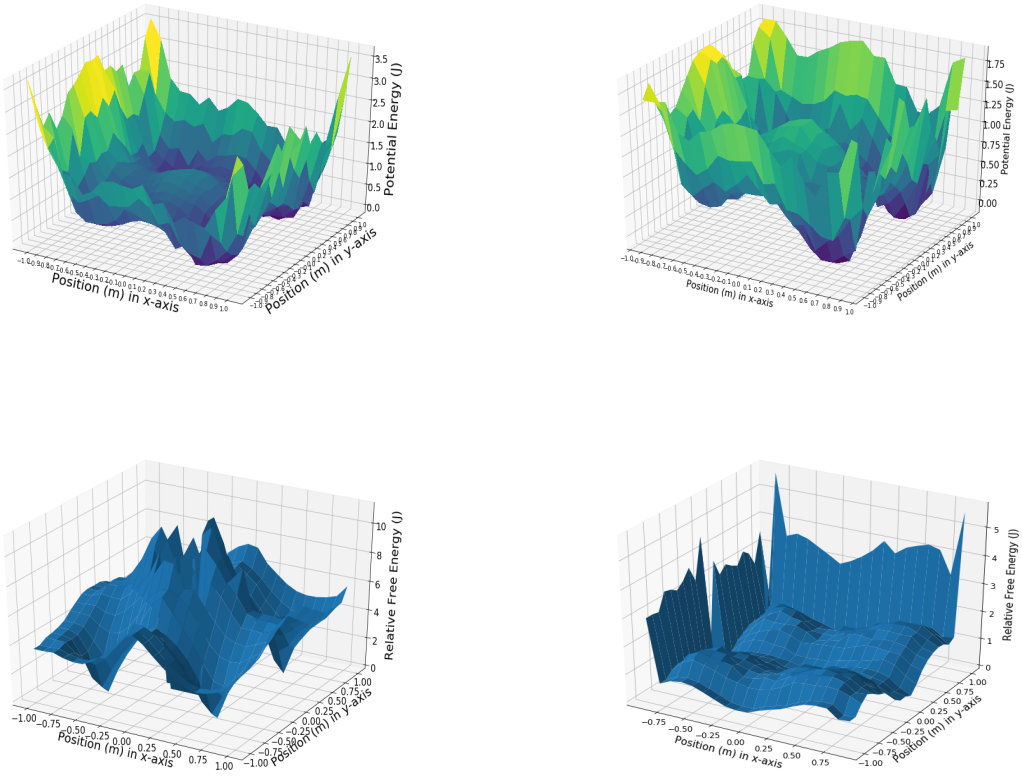

Figure 1.9: Potential energy (in green). Relative Free Energy plots (in blue)