Prof. Dr. Petra Imhof
Prof. Dr. Felix Höfling
Irtaza Hassan (irtaza06@zedat.fu-berlin.de)

Freie Universität Berlin
FB Mathematik und Informatik

Computational Molecular Physics **Winter 2017/18**

Problem set 6 — 7 December 2017

Problem 6.1 *k-means clustering*

The file *“datapoints.txt”* contains atomic cartesian coordinates xyz.
Use the k-means algorithm i.e.,

$$min \sum_{i=1}^{k} \sum_{x_n} \|x_n - c_k\|^2$$

where x_n are the data points and c_k are the cluster centres.

a) Implement the k-means clustering algorithm in your favourite programming language.

b) Use your code and run k-means on the provided dataset using the euclidean distance and different number of clusters i.e., $k = 5, 7, 10$ etc. How many number of clusters result in good clustering of the data?

c) Plot the clusters and the cluster centroids using different colors.

For implementation of the k-means clustering algorithm in Python and MATLAB see the following links:

[**Python**] *https://datasciencelab.wordpress.com/2013/12/12/clustering-with-k-means-in-python/*
[**MATLAB**] *https://de.mathworks.com/help/stats/kmeans.html*

Problem 6.2 *k-means clustering and implied time scales*

The file *“phiLpsiAdihedrals.txt”* contains $\Phi_{Leucine}$ and $\Psi_{Alanine}$ torsion angles values extracted from a trajectory, obtained from the molecular dynamics simulation of Alanine-Leucine. The first is time-step and second and third are torsion angle values.

a) Make a joint distribution plot (the ramachandran plot) of torsion angles $\Phi_{Leucine}$ and $\Psi_{Alanine}$ using the given data.

b) Run the k-means algorithm with k=4 clusters on the provided dataset using the euclidean distance. Also, repeat joint distribution plot using different color for each cluster.

c) Compute the transition matrix by counting the transitions between the clusters.

d) Choose 3 different lag-times i.e., $\tau = 2, 10, 50$ and again compute transition matrices.

e) Calculte implied time scales i.e., $t_i = \dfrac{\tau}{ln\lambda_i}$ using the eigenvalues of transition matrices estimated at different lag-times.

f) Make a plot of lag-times (x-axis) versus the corresponding implied time scale (y-axis). Do the implied time scales converge or not? Discuss your results.

Note: Please take into account the periodicity of the torsion angles while clustering and you can use any of the available utility for k-means clustering.

(please turn over)

Problem 6.3 *Chapman–Kolmogorov test*

The file *"dtraj.txt"* contains sequence of states, i.e., $[0, 1, 2, 3]$, extracted from the molecular dynamics trajectory of Alanine-Leucine after partitioning the selected torsion angles space.

a) Compute the transition matrix T by counting the transitions between the states for lag-time $\tau = 1$ units. Also compute the corresponding eigenvectors.

b) Again compute the transition matrix T for arbitrary lag-times τ i.e., $5, 10, 15, 20$ units.

c) Take n powers (i.e.,$= 5, 10, 15, 20$) of transition matrix T calcualted in a) i.e., $T(\tau)^n$, where n is the variable integer.

d) Apply transition matrices computed in b) and c) on $2^n d$ eigenvector computed in a) and compare. What do you observe?

Due date: ***14 December, 12 p.m.***