Mohammad Shahriar Hooshmand
Ph.D. Candidate
Department of Materials Science and Engineering
The Ohio State University

CSE 5523 HW5: EM

(a) Below are the train and test loglikelihood vs iterations for 4 different seeds (I trimmed above 30 iterations to 100 for the sake of clarity). Top figures show the loglikelihood and the bottom figures show the change in loglikelihood until convergence. The criterion for convergence is set based on the maximum number of iterations as well as the change in the previous and current iteration loglikelihood below 0.1%. Due to the unsupervised classification characteristic of this problem, it takes a few iterations for error to remains within the range of -0.1%-0.1%. Also, it is worth noting that in the test sets, the parameters shouldn't be updated (no M step). So, the inset in the lower figures show this fluctuation. My experiment shows that the iterations for training data until the convergence varies roughly from **5-20.** After training on the .train data, convergence is achieved in less than 5 iterations on the .test data.

(b) The results for 4 samples with different seeds are shown above. Figures below compare the loglikelihood on training and test data for 10 different seeds. As it is seen, using different seed in the randomized functions changes the values a bit, however, they are mostly in the same range. It can be observed that for some particular seeds (here seed=5), there might be an instance which generates outlier data which might even correspond to machine accuracy or the initial randomized values. So, it is always necessary to run a check test on various seeds before analyzing the results.

(c) Here, the script is written to extract the labels from true data. Then, a routine is added to EM class (EM.LLL) which basically splits the loglikelihoods of each data over different classes. Then, the maximum loglikelihood for each data will be the predicted cluster and is compared with the true label. The accuracy here is defined as the number of instances that label is predicted correctly over the total data points. As we see, again seed plays an important role in the results. In some instances, accuracy is predicted as high as 90% and as low as 10%, however most of the observation lies around 40% accuracy. I think this is fair due to the limited number of data points we have trained as well as the unsupervised classification character of problem.

(d) Figure below shows the loglikelihood for train and test sets over the range of different clusters. As we increase the clusters, loglikelihood increases. Also, loglikelihood for test sets are greater compared to the training set which is expected to occur. It is interesting that as the number of clusters increase, loglikelihood for test data increases as well which is fair as the clusters are pretty much guaranteed to have lower variance as the number of them increases. It is important to mention that the objective function of this algorithm is to find the maximum θ as follows:

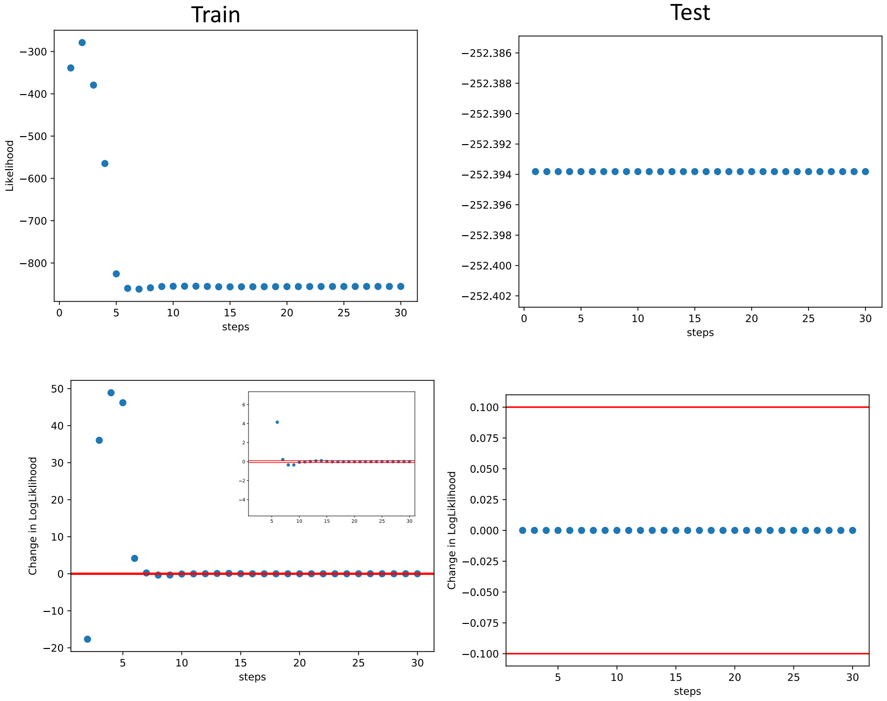
$$\text{argmax}_{\theta} \prod_j \sum_{i=1}^{k} P(y^j=i,x^j|\theta) = \sum_j \log \sum_{i=1}^{k} P(y^j=i,x^j|\theta)$$

As a result, we are indirectly choosing the $\theta$from Bayes rule which finds the maximum likelihood of the observed data. Knowing how many sources (clusters) needed requires prior knowledge (remember posterior is different with the likelihood). Also, I should emphasize that the prior we are considering in these notations has different meaning than the "prior knowledge on #centers". In EM, the "hidden" prior is provided by Gaussian distribution and the objective function follows the above equation based on MLE.

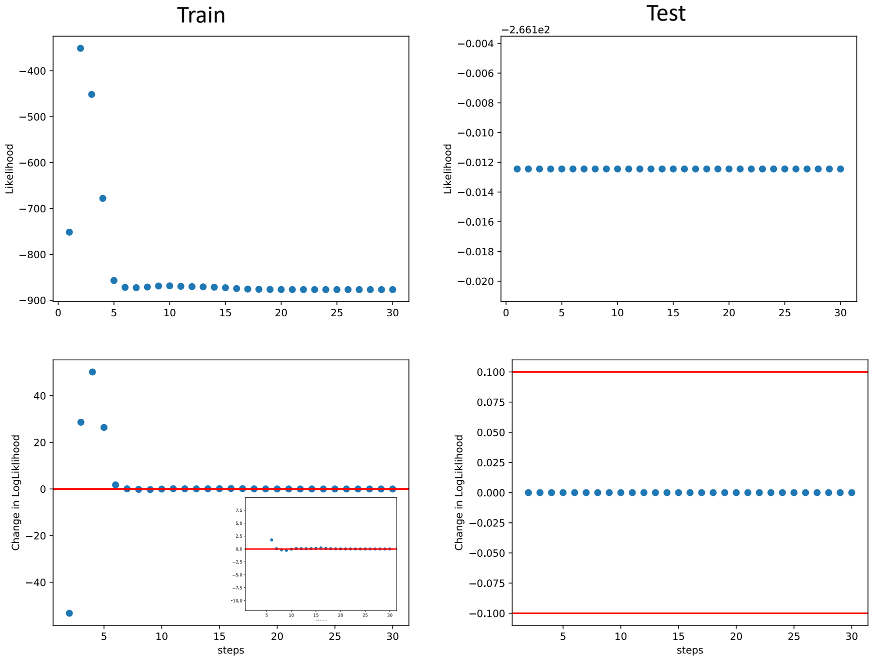
#Note: Each section of above problems are commented in the __main__ section of code. Each block can be uncommented to be functional.

Figures:

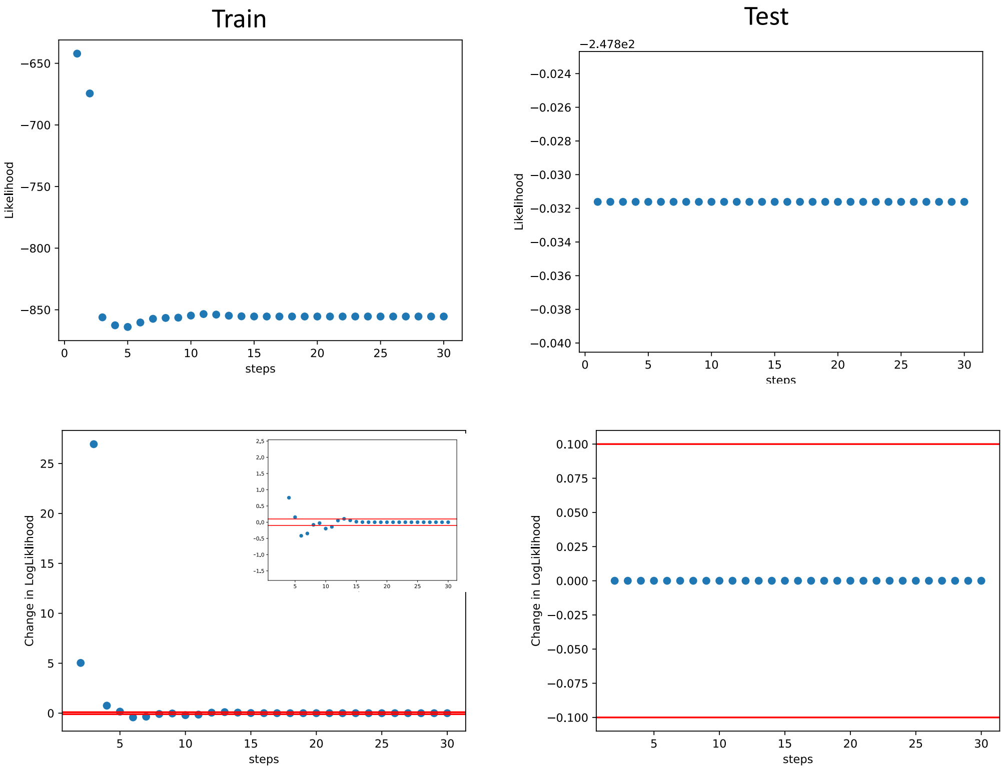
Seed Sample1



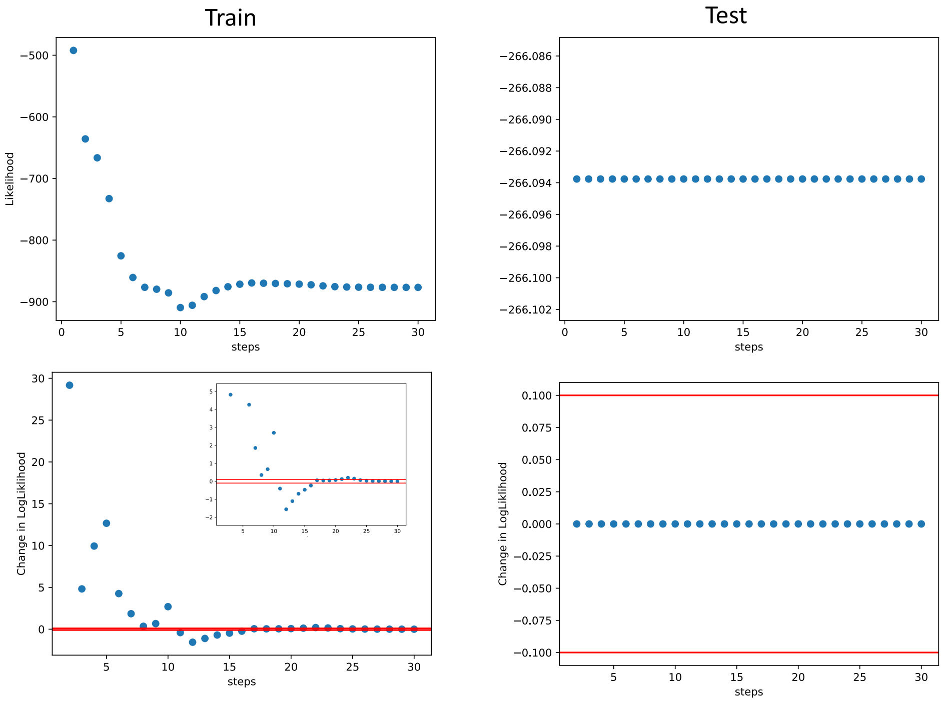
Seed Sample2

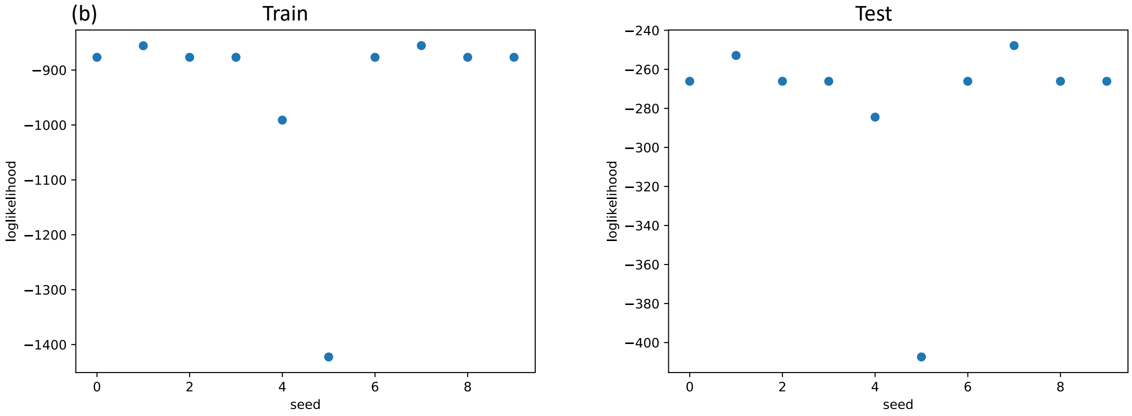


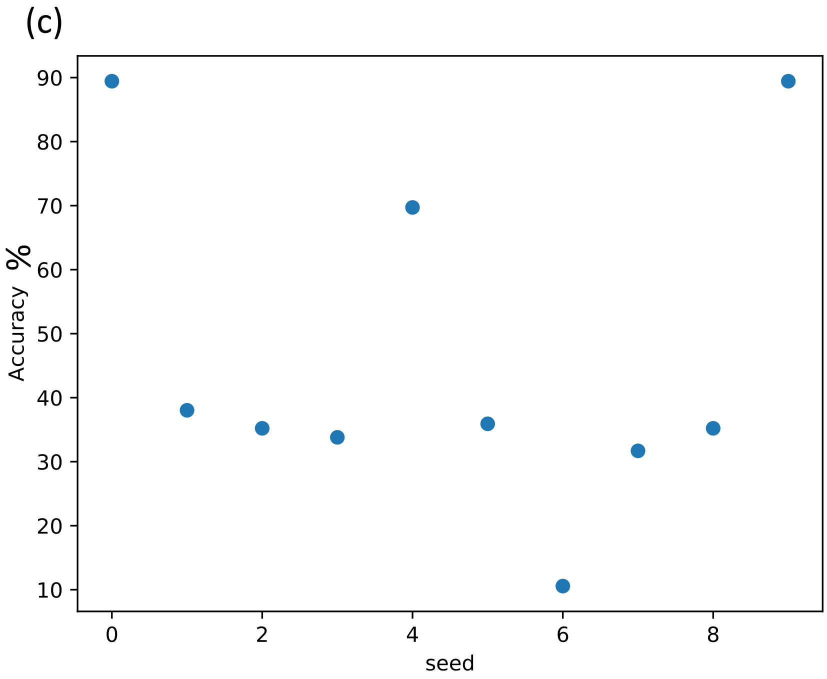
Seed Sample3

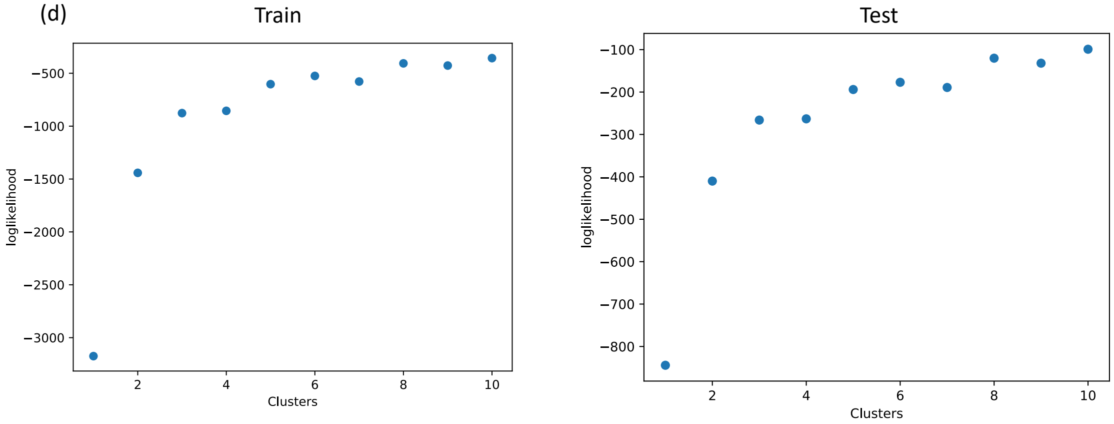


Seed Sample4




(b)
Train
−900
−1000
−1100
−1200
−1300
−1400
loglikelihood
0
2
4
6
8
seed
Test
−240
−260
−280
−300
−320
−340
−360
−380
−400
loglikelihood
0
2
4
6
8
seed


(c)
90
80
70
60
50
40
30
20
10
Accuracy %
0
2
4
6
8
seed


(d)
Train
−500
−1000
−1500
−2000
−2500
−3000
loglikelihood
2
4
6
8
10
Clusters
Test
−100
−200
−300
−400
−500
−600
−700
−800
loglikelihood
2
4
6
8
10
Clusters