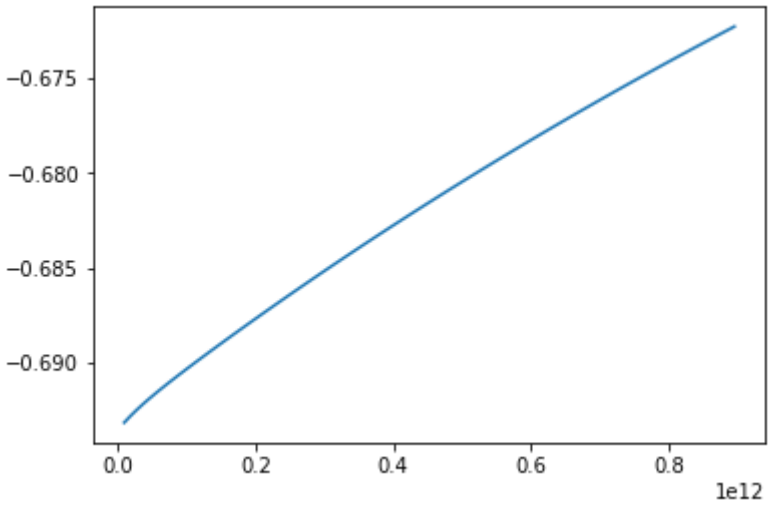
Machine Learning Report (Assignment 1)

Shashwat Shivam (2016CS10328)

1. Gradient Descent :-
   a. Constant Learning Rate = 0.001
      Iterations = 128
      Lambda = 0.01

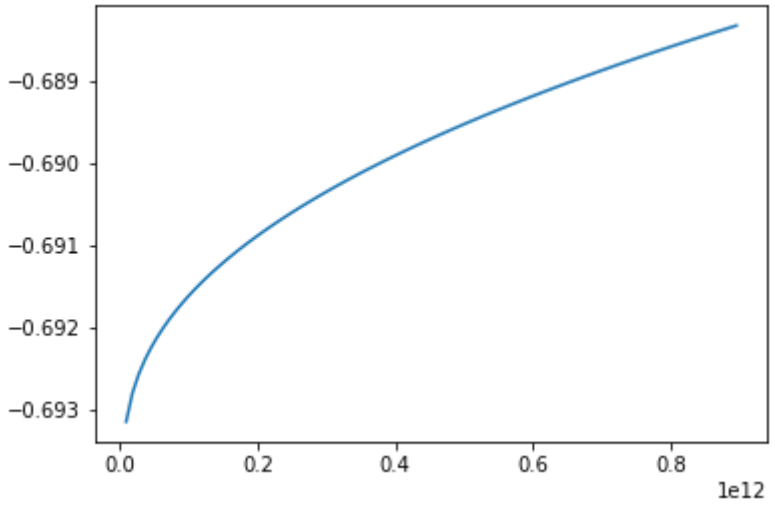
      Log Likelihood versus floating point iterations graph:-



      We can see that this graph is close to linear due to the constant learning rate.

   b. Adaptive Learning Rate:-
      Initial Value = 0.001
      Iterations = 100
      Lambda = 0.01

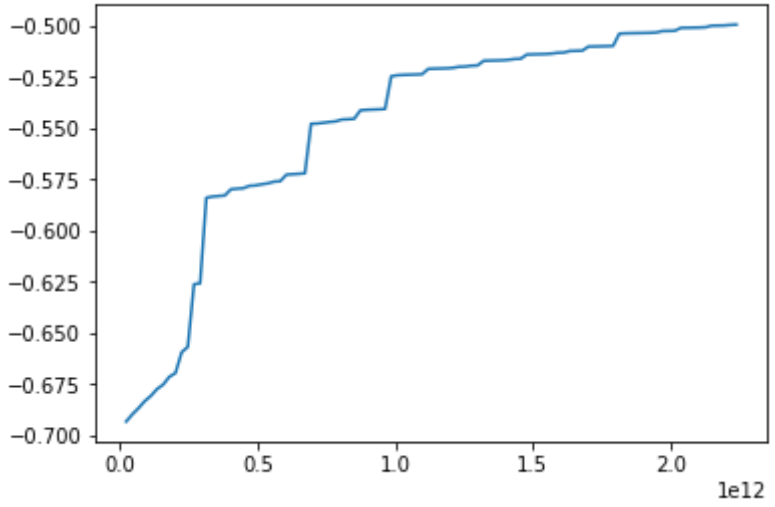
      Log Likelihood versus floating point iterations graph:-



We can see that due to the decreasing learning rate we get a better (faster) drop in log likelihood for the same number of iterations.

c. Binary Search optimal learning rate:-
Iterations = 100
Lambda = 0.01

Log Likelihood versus floating point iterations graph:-

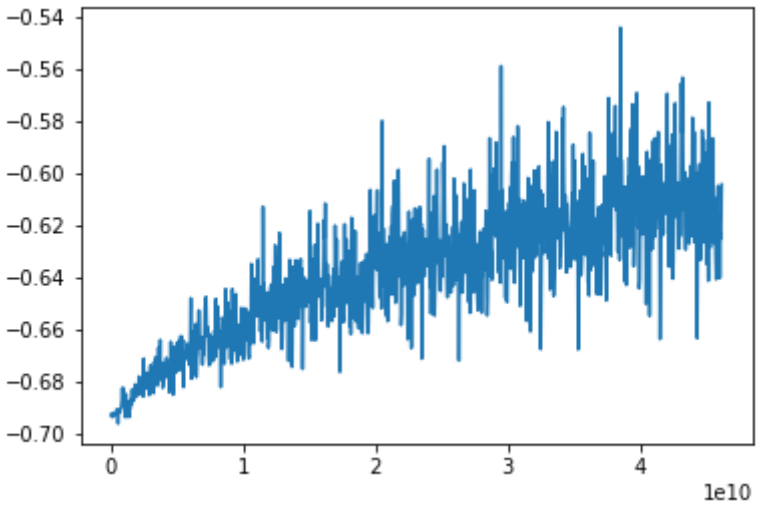


We can see that for the same values on the x-axis the log likelihood decreases much faster if we search for optimal learning rate value. This method performs better than both the above methods by a wide margin.

2. Stochastic Gradient Descent:-
   a. Constant Learning Rate = 0.01
   Iterations = 100
   Batch Size = 128

Lambda = 0.01

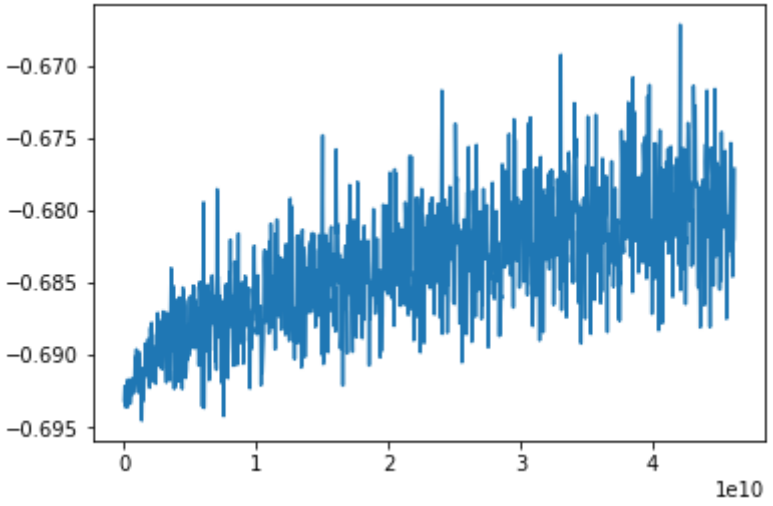
Log Likelihood versus floating point iterations graph:-



We can see from this graph that there is constant variation in log likelihood. This variation comes due to the small batch size which is not able to denote a general trend. On comparing this with Gradient descent we can see that in general a higher likelihood is obtained with much less number of floating point operations. This can show the efficiency of stochastic gradient descent.

b. Adaptive Learning Rate:-
Initial Learning Rate = 0.01
Iterations = 100
Batch Size = 128
Lambda = 0.01

Log Likelihood versus floating point iterations graph:-



This model performs worse than the previous model as the weight

given to each batch is different due to the constantly changing learning rate. This model shows a higher variance and change in log likelihood during training. The results are also worse than other models.

c. Binary search learning rate:-
   This model could not be successfully created as the binary search on a small batch can result in a learning rate which takes the model away from the general trend. As a result of this many of the log-likelihood values in the training process were out of bounds and could not be plotted.

3. I experimented with 2 stopping criterion:-
   a. Fixed number of iterations:- This method is good because it gives a good estimate of the time it will take to train a particular model. Although depending on the number of iterations specified the model can be good or bad. The accuracy of the model generally increases with increasing number of iterations (rate of change may decrease).
   b. The rate of change of error/ log-likelihood:- This method is good because we can directly train a model which provides at least a certain level of accuracy. The problem with this method is that the algorithm may take a lot of time to reach the required error threshold. This may also result in overfitting due to a huge number of iterations if the rate of change when to stop training is set very low.

      Based on experiments in part a and b Gradient Descent with the binary search of learning rate leads to the best convergence of the model. If the number of floating point operations has to kept less then Stochastic Gradient Descent with a constant learning rate can be used.