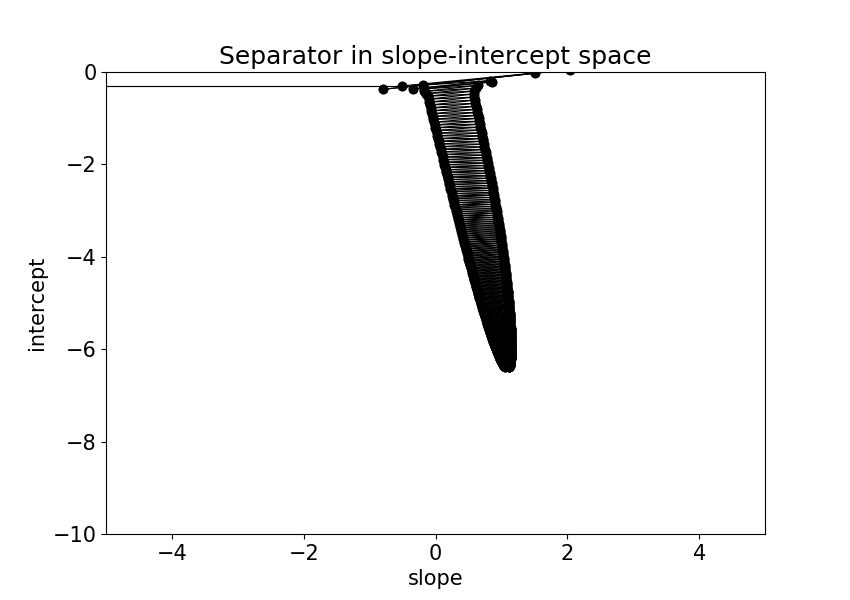
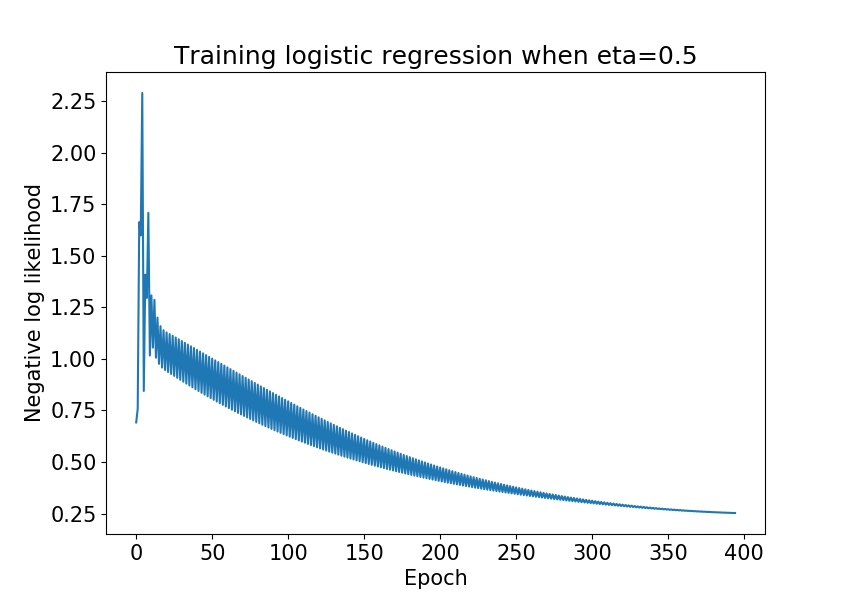
Question 5

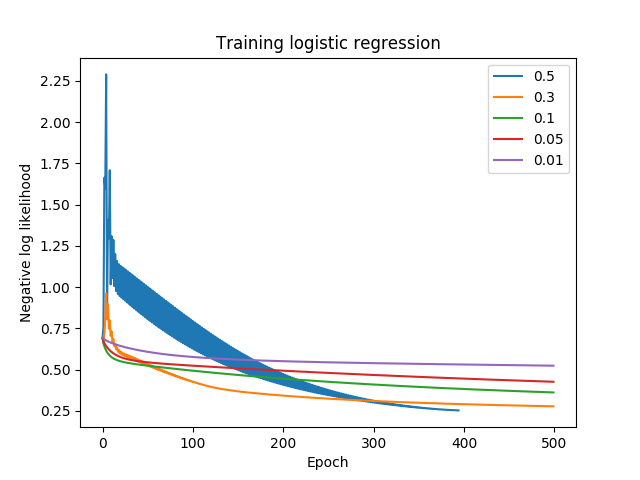
1)





The oscillating of the error curve could be caused by the relatively large learning rate eta=0.5. The error decrease in the direction of the gradient but if the learning rate is too large we could step over the local minimum and jump to some instances with even higher error.

2)



The relatively small learning rate 0.3, 0.1, 0.05, 0.01 gives less oscillating curve since the step we move each time in gradient decent now is smaller.

The relatively large 0.5 learning rate gave us the lowest error overall and it achieves the tolerance level most quickly, in other words before 500 epochs.

The 0.1 learning rate gave us the low error most quickly as you can see on the graph that green line is the steepest before 100 epochs.

So I think there’s no fixed answer for the learning rate choice in this case. It all depends on the experiment setting.

3)

I have talked with TA about this question since I noticed that shuffle made a big difference for the result at different place (shuffle for each epoch/eta or shuffle at the beginning). The TA explained to me that usually shuffle doesn’t affect the model a lot and the most reasonable place to shuffle is for each epoch since it increases the possibility to find the right w. But for our case