

Assignment 2

Anita Bahmanyar

November 14, 2016

Question 3

3.1

3.2

NN

This part of the assignment is asking for running the neural network code with different values of hyper parameters. The first part is to fix all the hyper parameters and change epsilon values from 0.001 to 1.0 for 5 different epsilon values. Figure 1 shows the cross entropy and accuracy of validation set for all 5 different epsilon values which I chose to be 0.001, 0.01, 0.1, 0.5 and 1.0.

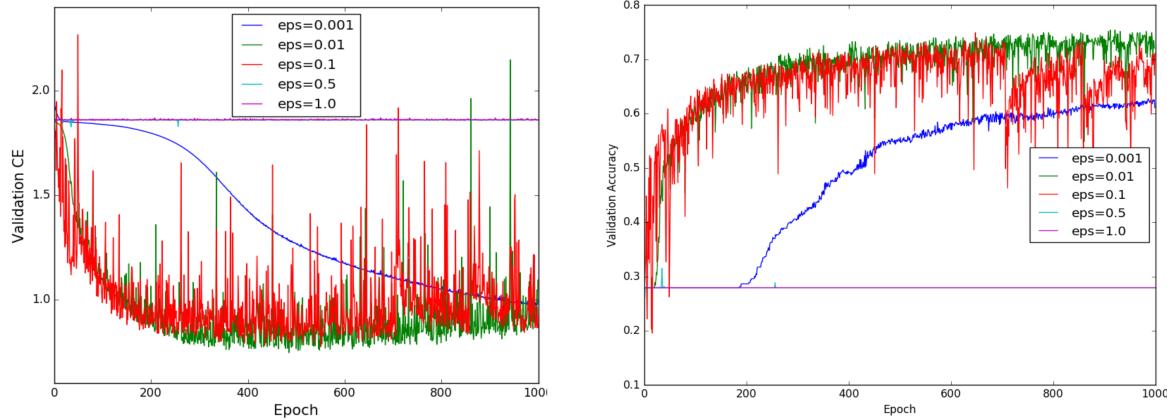


Figure 1:

What we conclude is that if epsilon is very small do instance in the case of 0.001, it takes very long for the validation to converge which means longer computation time. Also, if epsilon value is very large, for instance 0.5 and 1.0, then the accuracy curve would not be improving and it stays around 0.3 since we take large steps and we miss the global optimum value. I found that epsilon=0.01 and epsilon=0.1 have similar values and since the initialization is random, it is hard to say which is better since the accuracy of using

these values are pretty similar and for each run they differ due to the randomness. The value of epsilon I choose is **epsilon=0.01** to do the other runs and I keep it 0.01 which changing other parameters.

In Figures 2 and 4 I show few different plots that are for the same epsilon values but they are separate plots for each epsilon so that we can see the training curve as well.

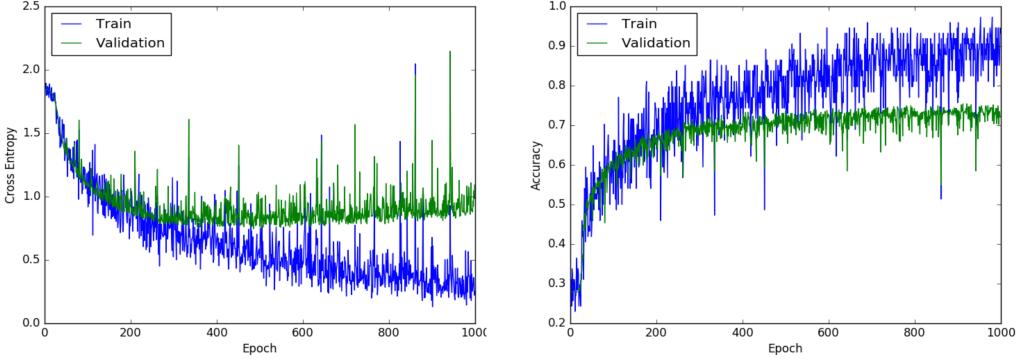


Figure 2: **epsilon=0.01**, CE on the left and accuracy on the right for both validation and training sets.

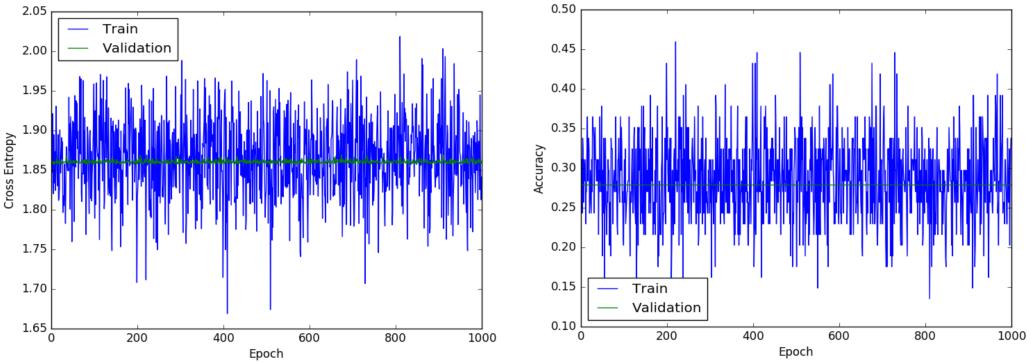


Figure 3: **epsilon=1.0**, CE on the left and accuracy on the right for both validation and training sets.

We also see that the accuracy of training set is always better than validation set which makes sense since we are training the model based on the training set. This is also true for cross-entropy where the training set always has smaller CE compared to the validation set.

In this part, I keep epsilon the same (0.01) and only change momentum value to be 0.0, 0.5 and 0.9. As it is shown in Figures ?? below, we see that momentum=0.9 does not return good results and we conclude it is too large. The results of momentum=0.0 and momentum=0.5 are similar except for the cross-entropy where for the case of momentum=0.5, the validation cross-entropy starts to go up at a much earlier time compared to momentum=0.0. When the validation CE starts increasing instead of decreasing that is the point where training further does not help which means it does not matter if we train more except that we spend more computational time, so the momentum value that I think is better in this case is **momentum=0.0** and I will use this value for the next part of the question.

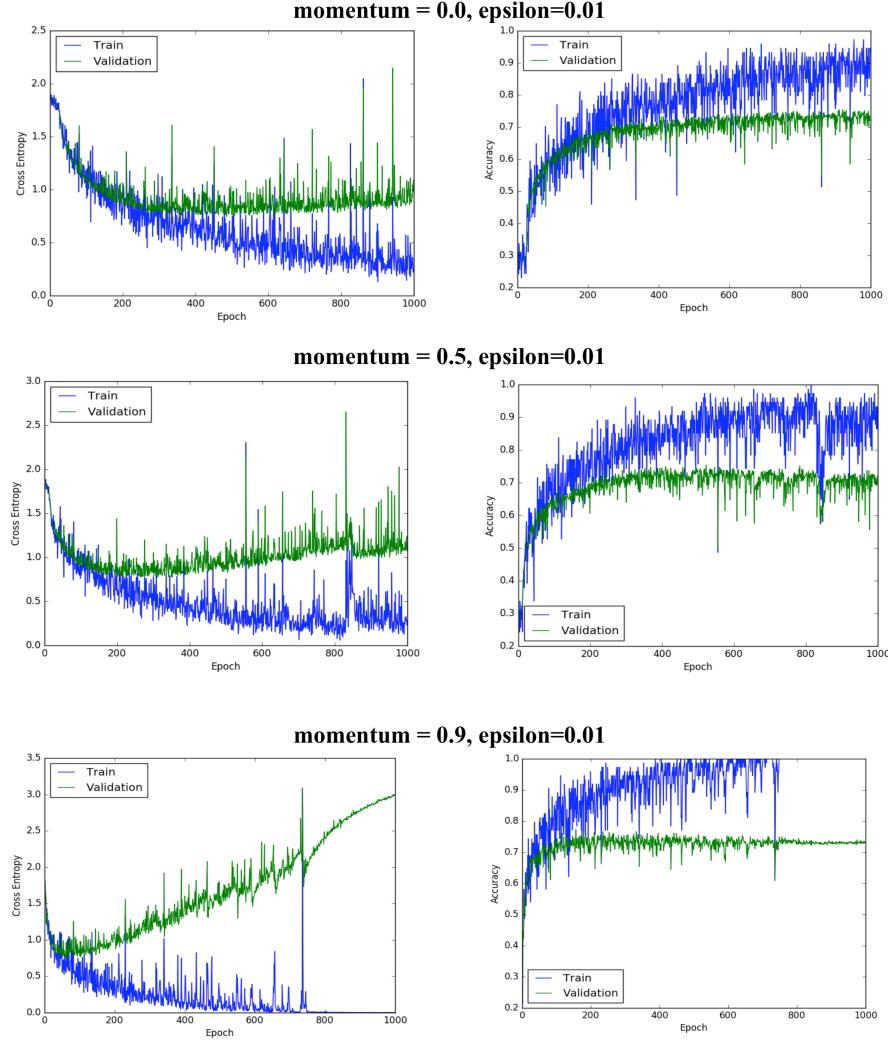


Figure 4: **epsilon=1.0**, CE on the left and accuracy on the right for both validation and training sets for three different momentum values while keeping **epsilon=0.01**. The momentum values are 0.0, 0.5 and 0.9 from top to the bottom.

Smaller batch sizes take longer to run due to having more steps to run since it is defined in the code that

```
num_steps = int(np.ceil(num_train_cases / batch_size))
```

so smaller batch size means more steps.

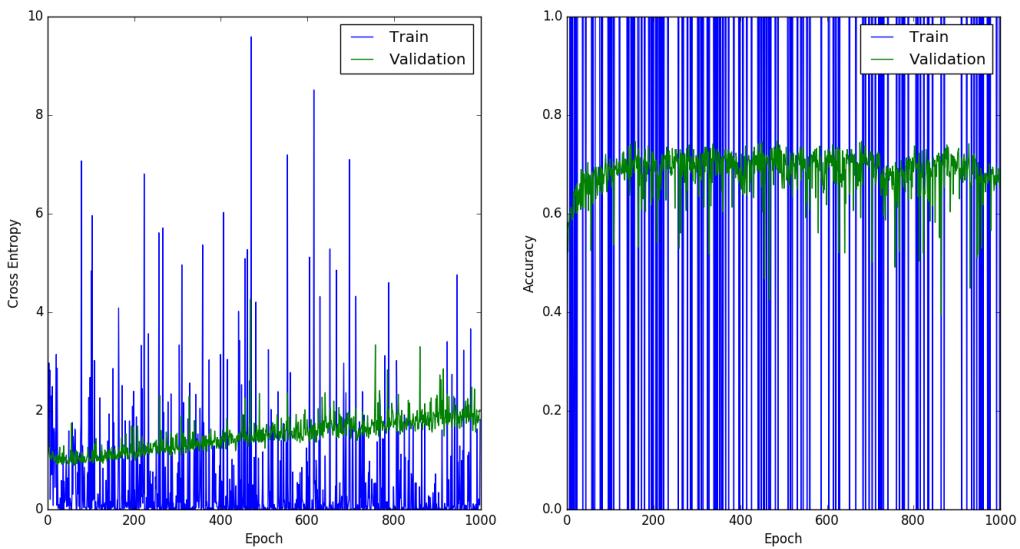


Figure 5: $\text{epsilon}=0.01$, $\text{momentum}=0.0$, $\text{batch size}=1$

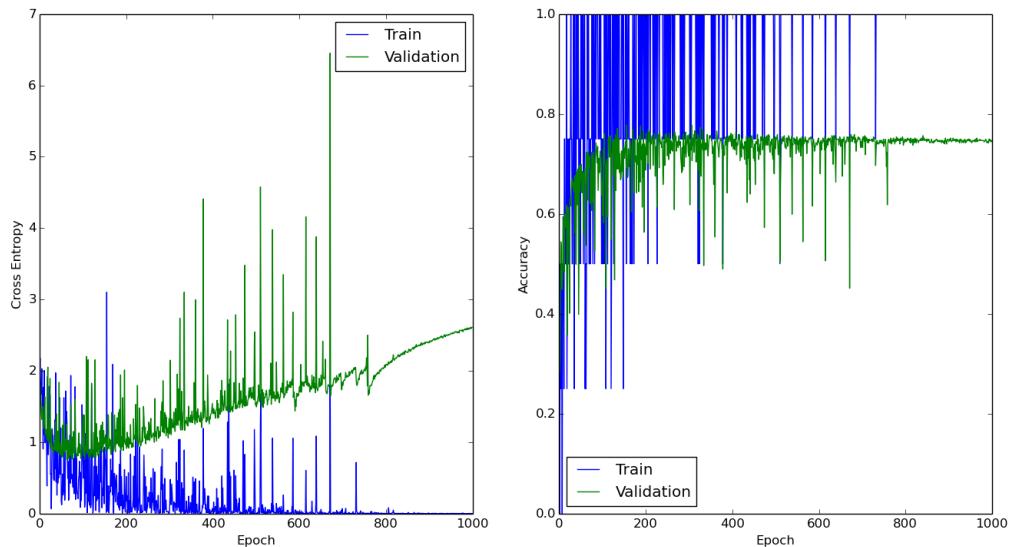


Figure 6: $\text{epsilon}=0.01$, $\text{momentum}=0.0$, $\text{batch size}=10$

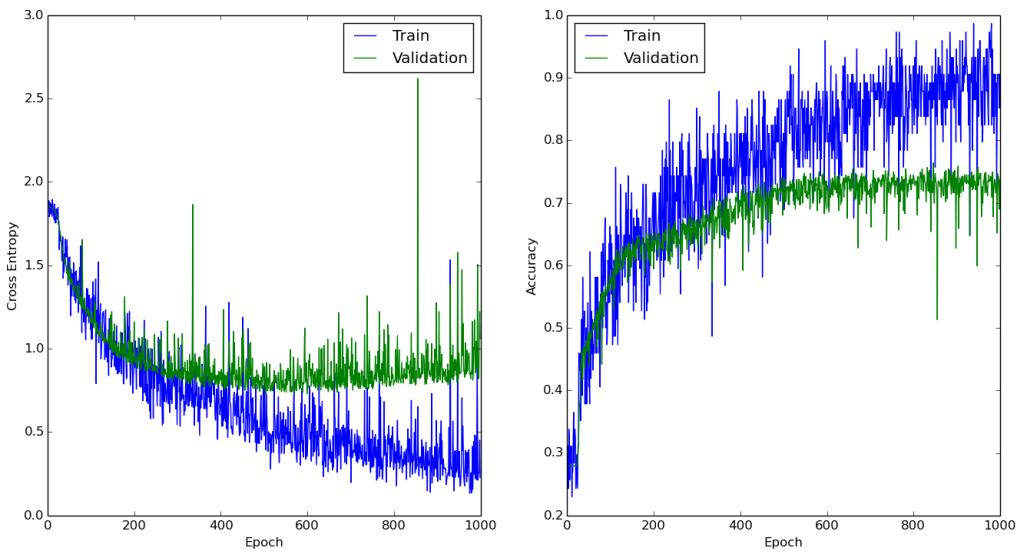


Figure 7: $\text{epsilon}=0.01$, $\text{momentum}=0.0$, $\text{batch size}=100$

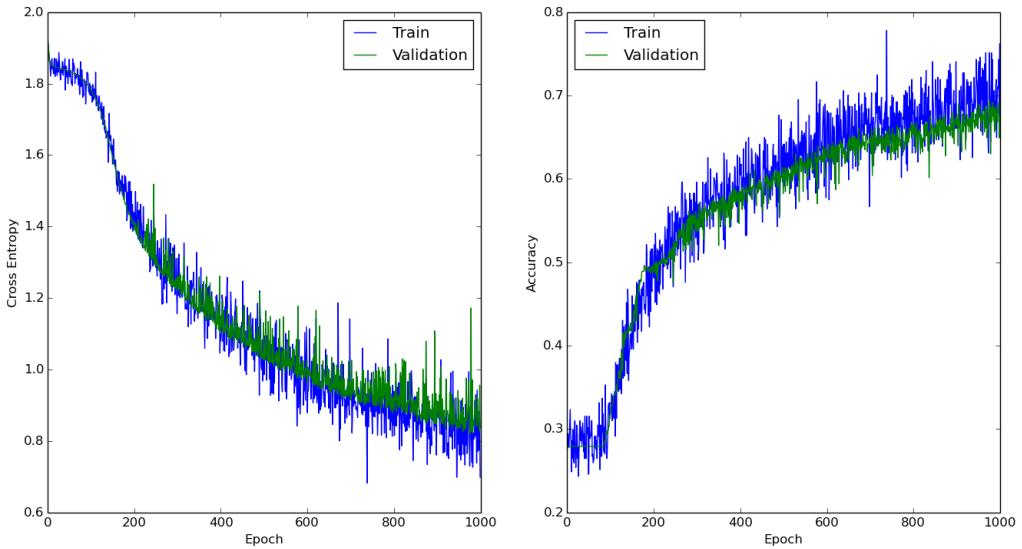


Figure 8: $\text{epsilon}=0.01$, $\text{momentum}=0.0$, $\text{batch size}=500$

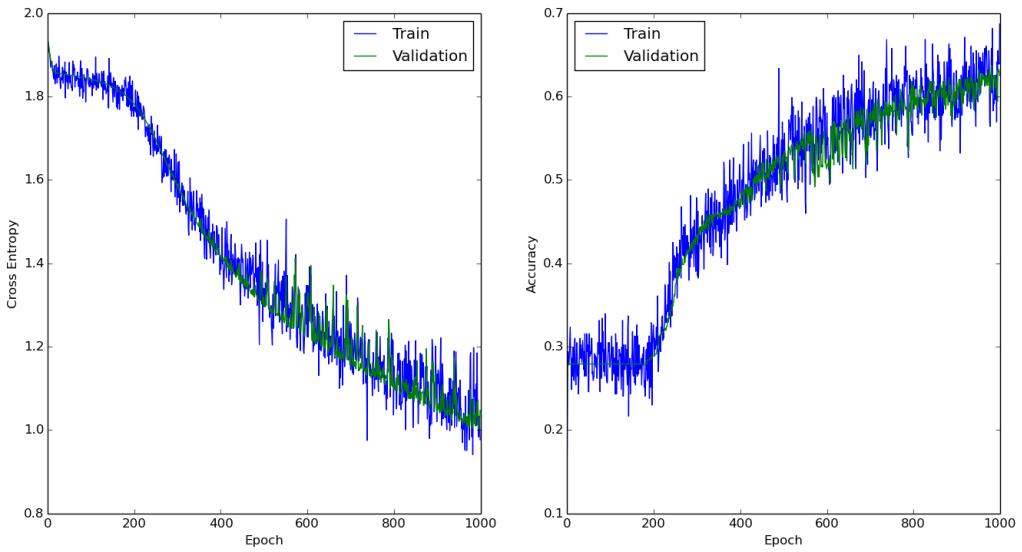


Figure 9: **epsilon=0.01, momentum=0.0, batch size=1000**

What I conclude is that the smaller batch size results in worse accuracy and bad cross-entropies. The higher batch size values result in higher accuracies and lower cross-entropies up to batch size=100. Also, for the case of batch size=1000, accuracy of validation set is very close to that of training set as well their cross entropies. This is because the training accuracies become lower. Also, for batch size values higher than 100, the accuracies converge slower. So the best value for **batch size=100**.

CNN

I ran CNN code for five different epsilon values and then compared their accuracies and their cross-entropies. Below are the results for each run.

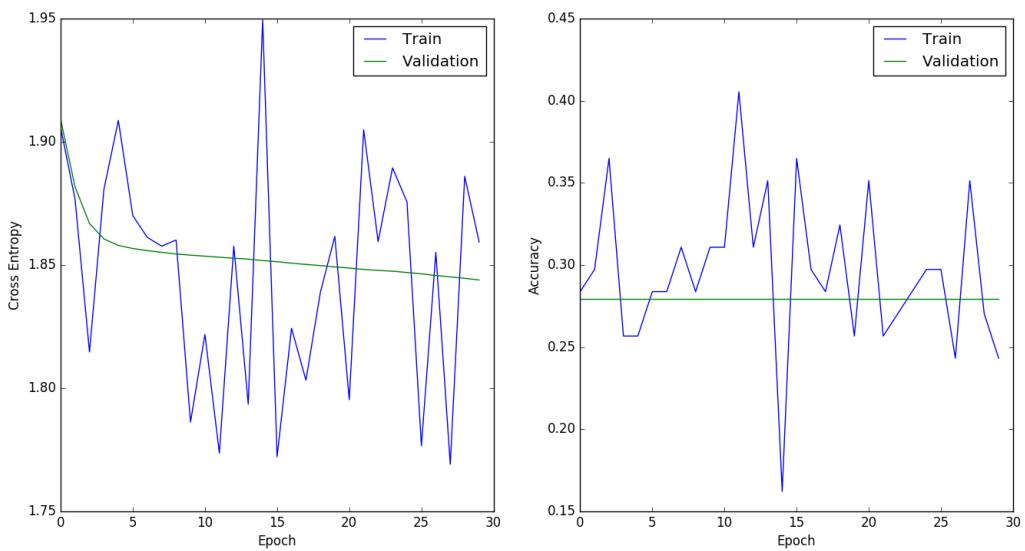


Figure 10: $\epsilon=0.001$, $\text{momentum}=0.0$, $\text{batch size}=100$

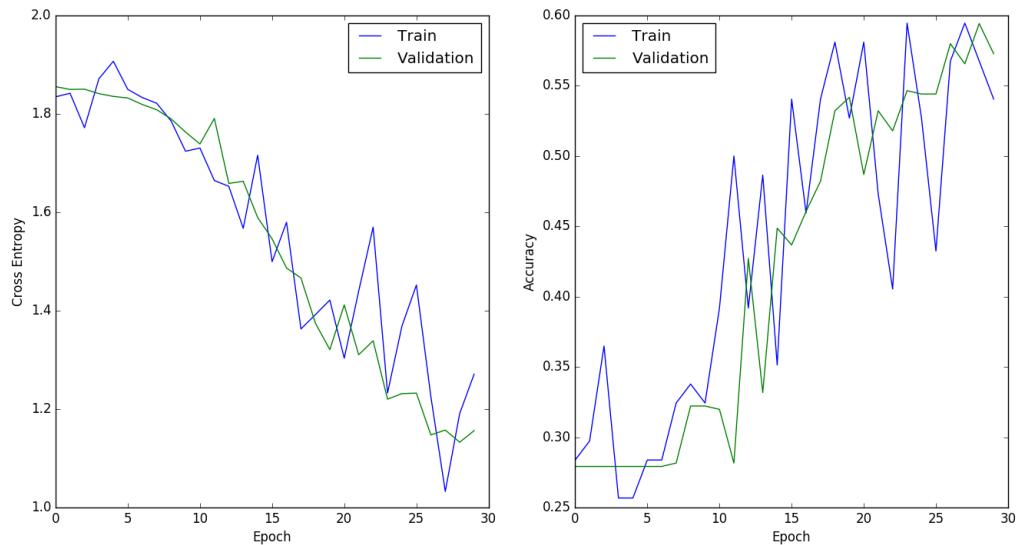


Figure 11: $\epsilon=0.01$, $\text{momentum}=0.0$, $\text{batch size}=100$

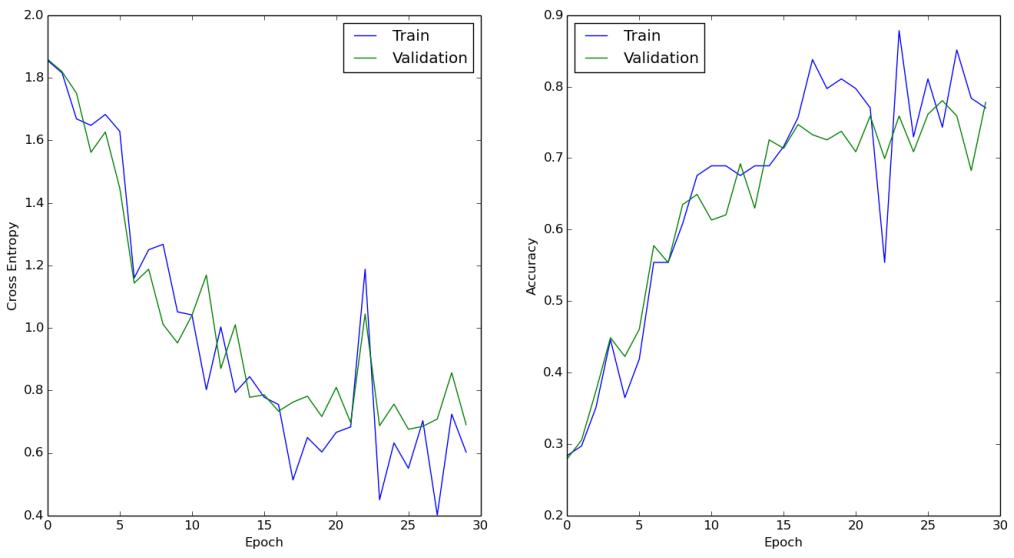


Figure 12: $\text{epsilon}=0.1$, $\text{momentum}=0.0$, $\text{batch size}=100$

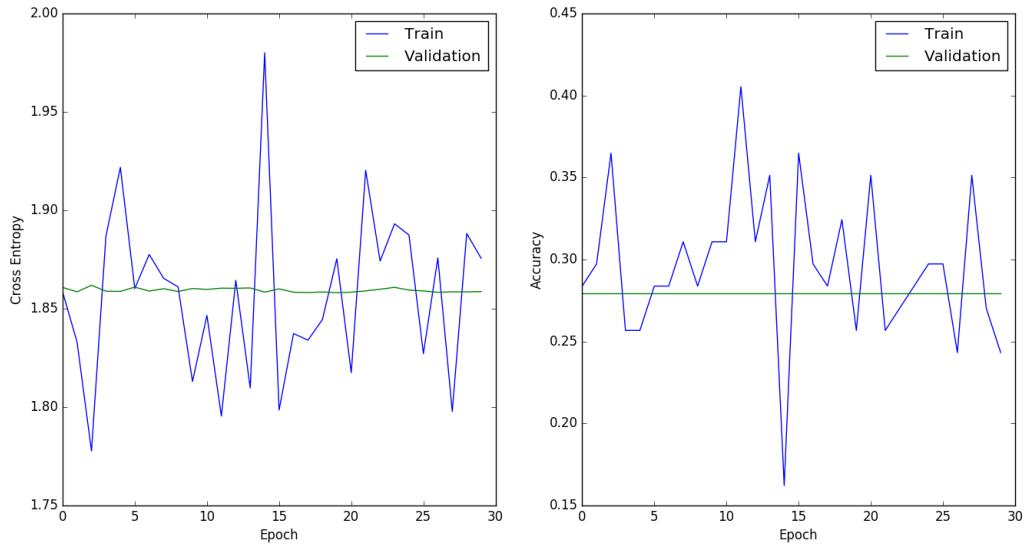


Figure 13: $\text{epsilon}=0.5$, $\text{momentum}=0.0$, $\text{batch size}=100$

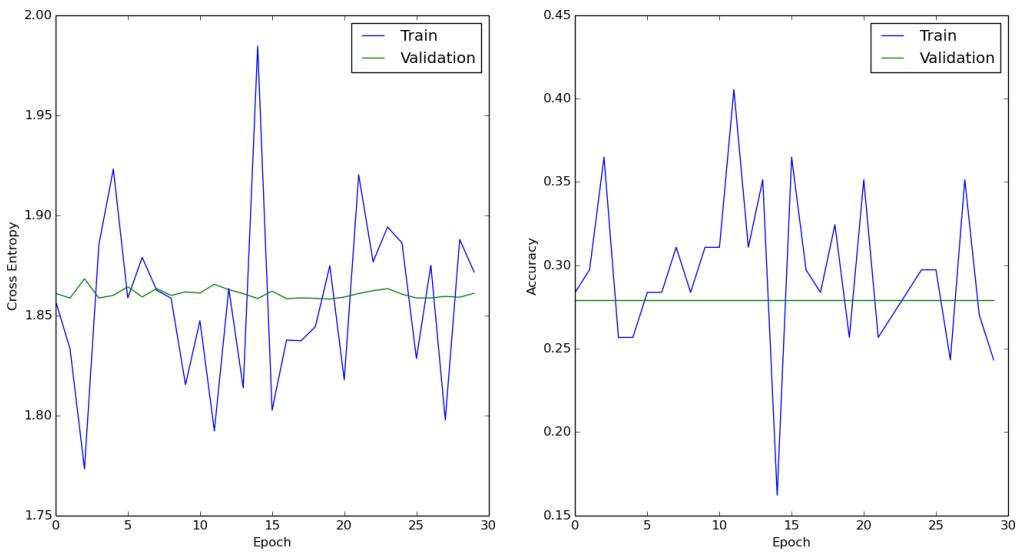


Figure 14: **epsilon=1.0, momentum=0.0, batch size=100**

Based on the figures above, **eps = 0.1** returns the best results using CNN method.

Fixing **eps = 0.1**, I ran CNN code using three different momentum values and the results are shown in figures below:

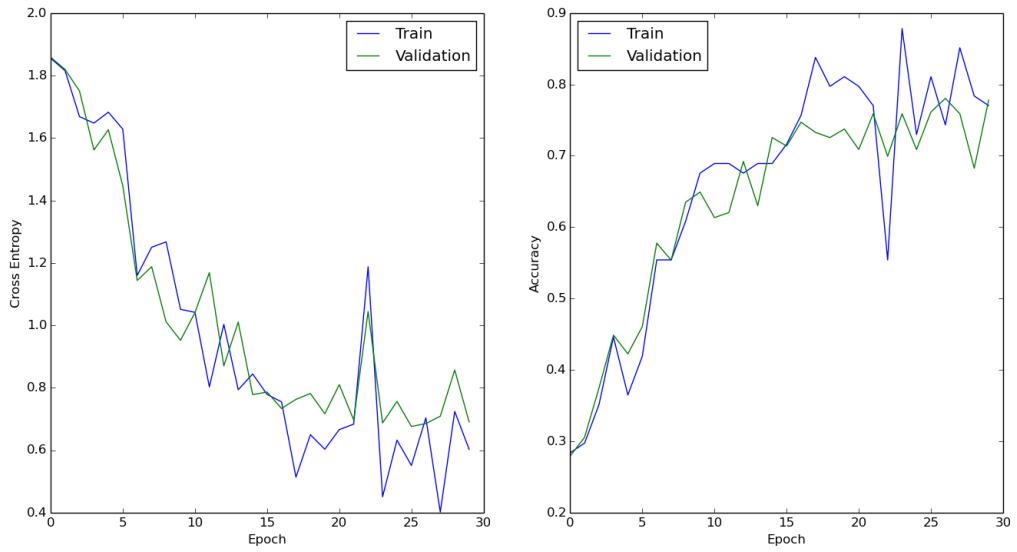


Figure 15: **epsilon=0.1, momentum=0.0, batch size=100**

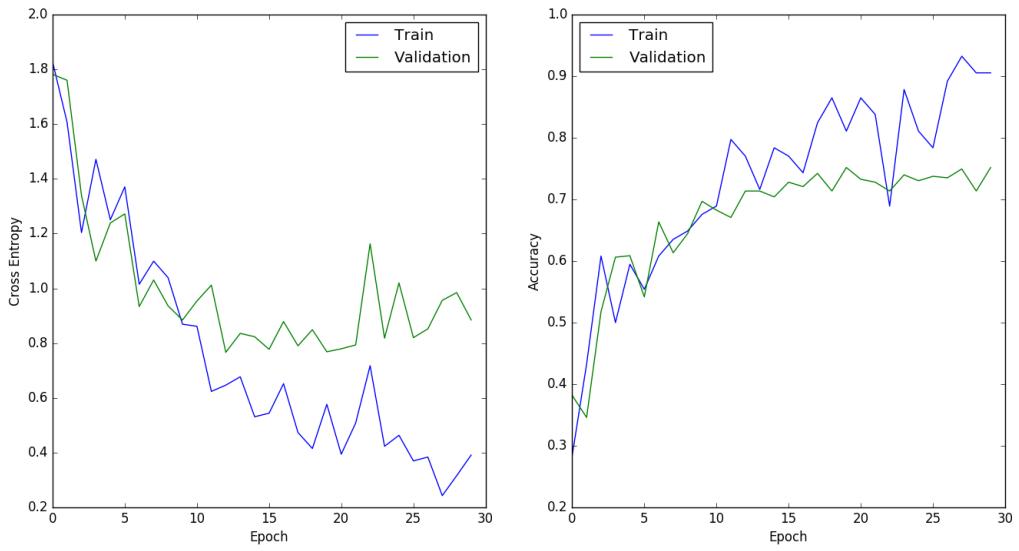


Figure 16: $\text{epsilon}=0.1$, $\text{momentum}=0.5$, $\text{batch size}=100$

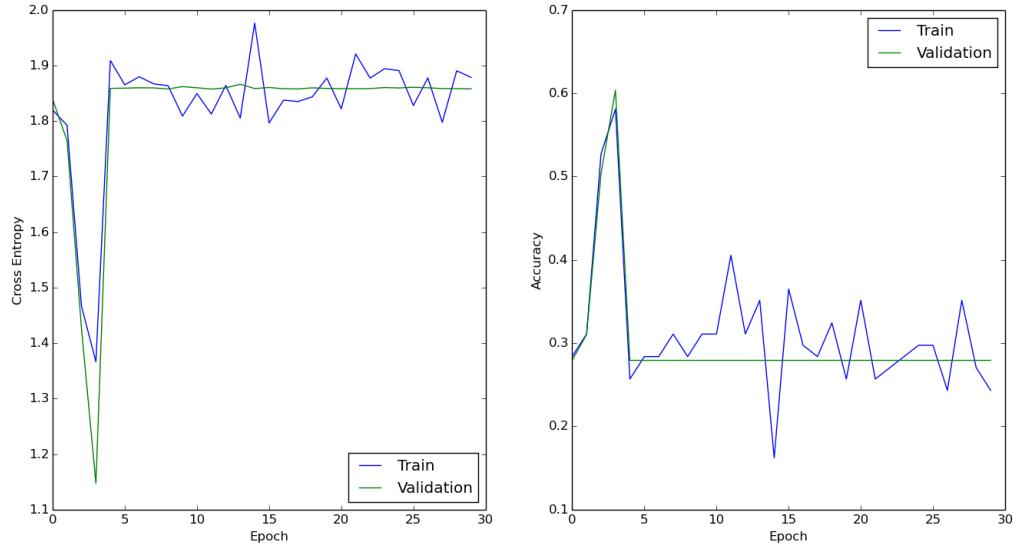


Figure 17: $\text{epsilon}=0.1$, $\text{momentum}=0.9$, $\text{batch size}=100$

As we see from the figures above, the best model is a result of using **momentum=0.0**. Using momentum=0.9 results in very bad accuracy and cross-entropy and using momentum=0.5 has lower accuracy and higher CE than when the momentum is 0.

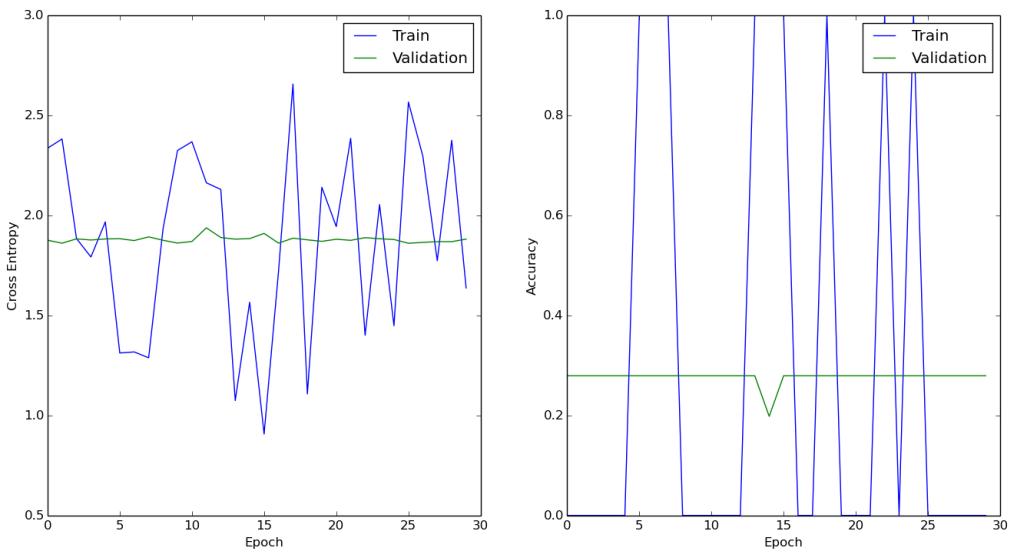


Figure 18: $\text{epsilon}=0.1$, $\text{momentum}=0.0$, $\text{batch size}=1$

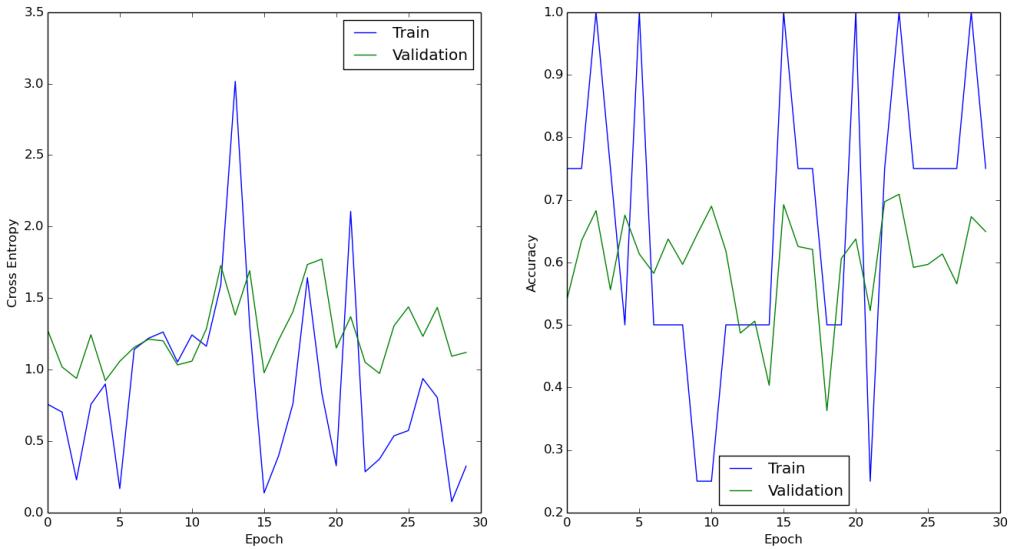


Figure 19: $\text{epsilon}=0.1$, $\text{momentum}=0.0$, $\text{batch size}=10$

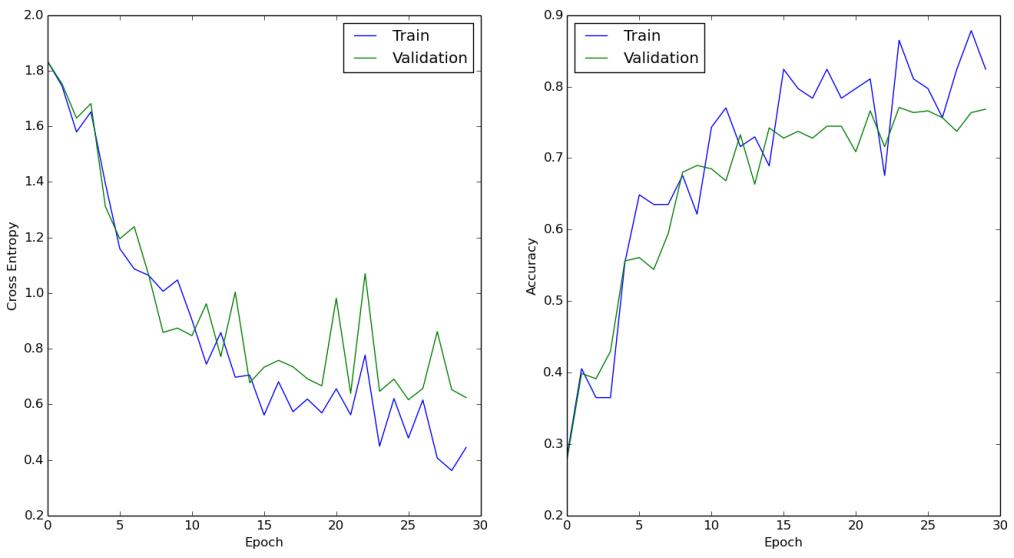


Figure 20: $\text{epsilon}=0.1$, $\text{momentum}=0.0$, $\text{batch size}=100$

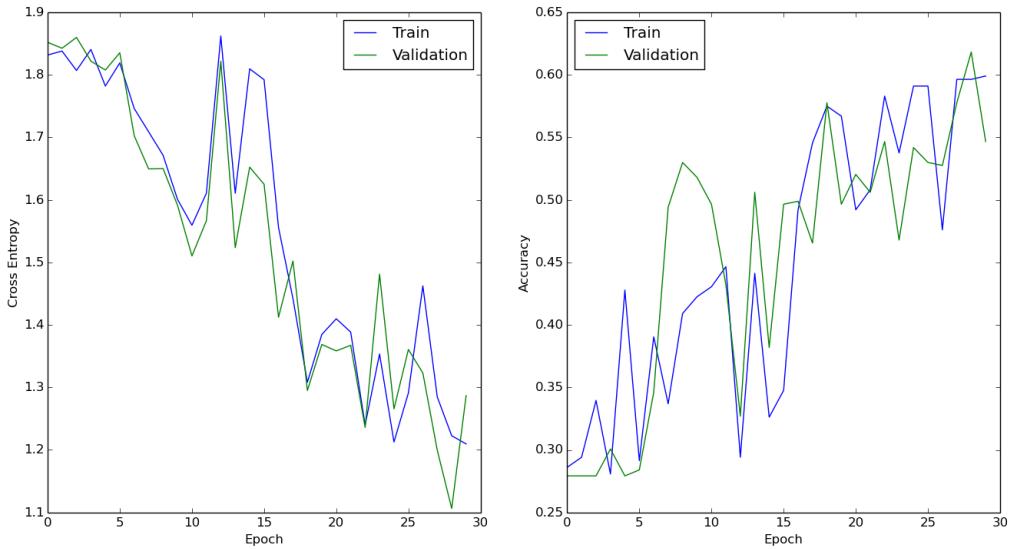


Figure 21: $\text{epsilon}=0.1$, $\text{momentum}=0.0$, $\text{batch size}=500$

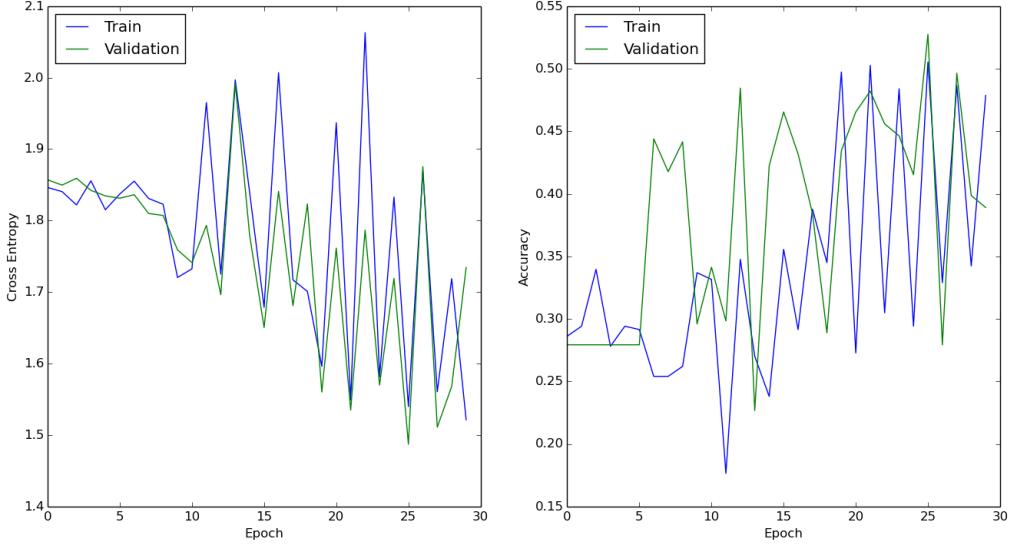


Figure 22: **epsilon=0.1, momentum=0.0, batch size=1000**

Looking at the figures above, we see that batch size=100 has the best results since the accuracy of the validation set is higher and it has lower cross entropy. Batch size=500 and batch size=1000 have lower accuracies around 0.45 -0.6 which is lower than 0.8 for the 100 batch size. Accuracy for batch size = 10 is about 0.6 and stays the same the whole time with no improvement and batch size=1 generates accuracy of as most 0.3 and does not improve either and has high cross-entropy above 1.5.

3.3

3.4

3.5

Question 4

4.2

In the code provided (mogEM.py), inside function mogEM there are a few lines that include "randConst" value.

```
p  = randConst + np.random.rand(K, 1)
p  = p / np.sum(p)    # mixing coefficients
mu = mn + np.random.randn(N, K) * (np.sqrt(vr) / randConst)
```

The first line will show π_k values which are the mixing coefficients in the Gaussian mixture method. The mixing coefficients will be dominated by random values if randConst is small and it will be dominated by the value of randConst if it is large. In the third line, as we increase randConst, " $(\text{np.sqrt}(\text{vr}) / \text{randConst})$ " decreases and as we decrease the randConst value, this part of "mu" expression increases. This means that

as we increase the randConst value, the "mu" values will be dominated by the mean values since the second part of "mu" expression will be very small.

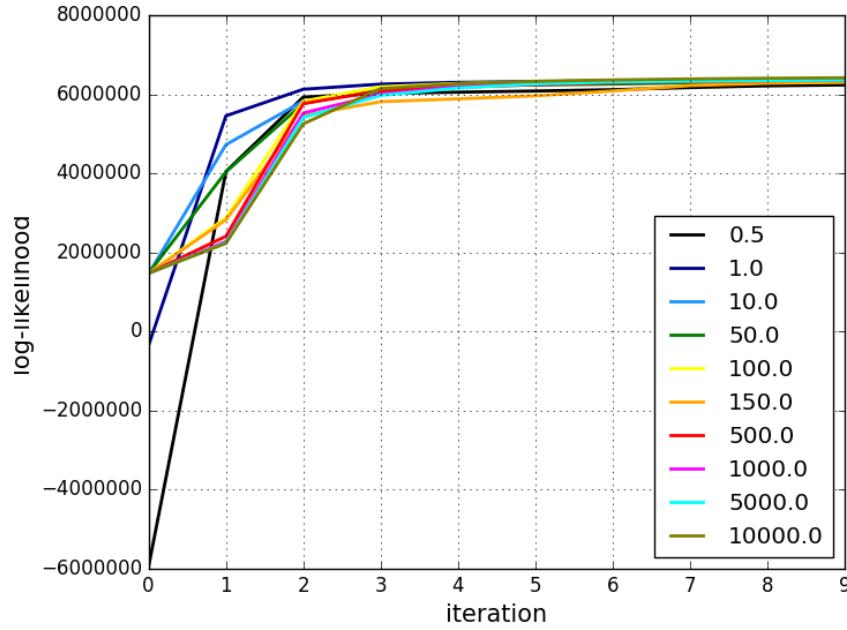


Figure 23: Log-likelihood as a function of number of iterations for different values of randConst parameter.

As we see in Figure 23, randConst=1.0 converges faster to the similar log-likelihood values compared to the other randConst values. So the model I choose has randConst=1.0. Below are variance and mean of the images shown for randConst = 1.0:

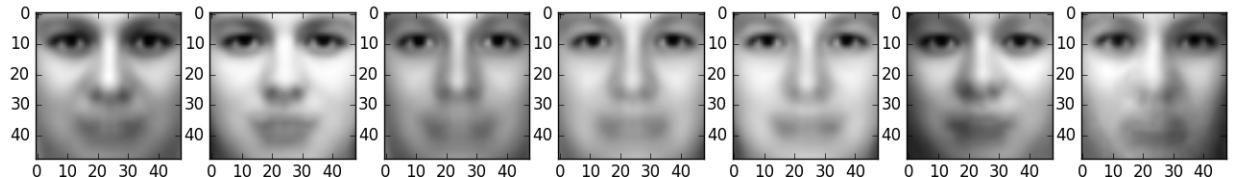


Figure 24: Mean of the images- Is is blurry because it is average.

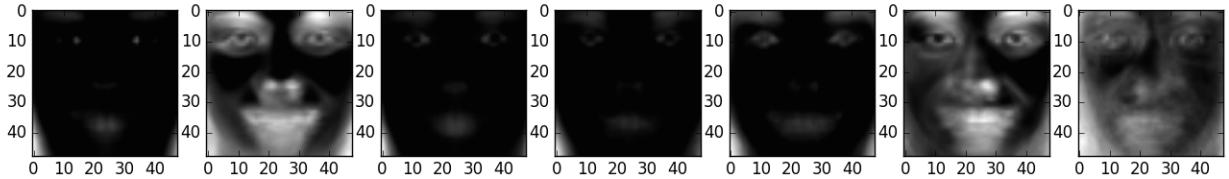


Figure 25: Variance of the images-black means lower variance and white means higher variance.

We see that most of these images look the same to us even though they are not and this is because there are small variations in the images. So variances are the most helpful ones in this case to see the variation between images. Also, black in the variance images mean lower variance and whiter means higher variance. Areas with lower variance are better so it is better if we see a lot of black areas in the variance images. Lower variance means better classifications so the variance images with more black area are better classifiers.

Below in Figure 26, the values of Gaussian mixture coefficients are given vs. cluster number.

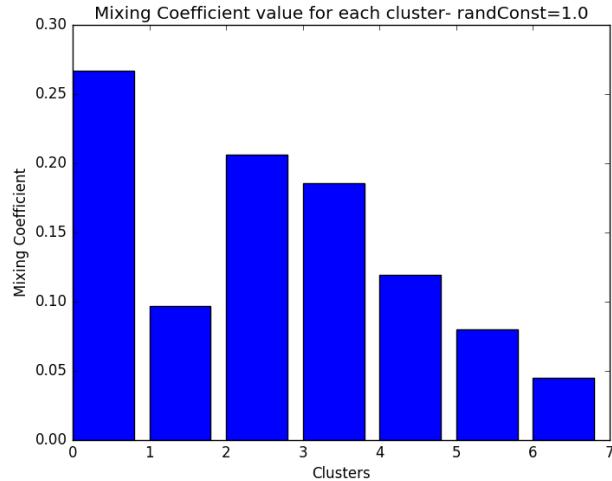


Figure 26: Value of mixture of coefficients for each cluster.

4.3

Initializing the means using K-means makes the code converge much faster than just using random values for the means. This can be seen in Figure 29 below.

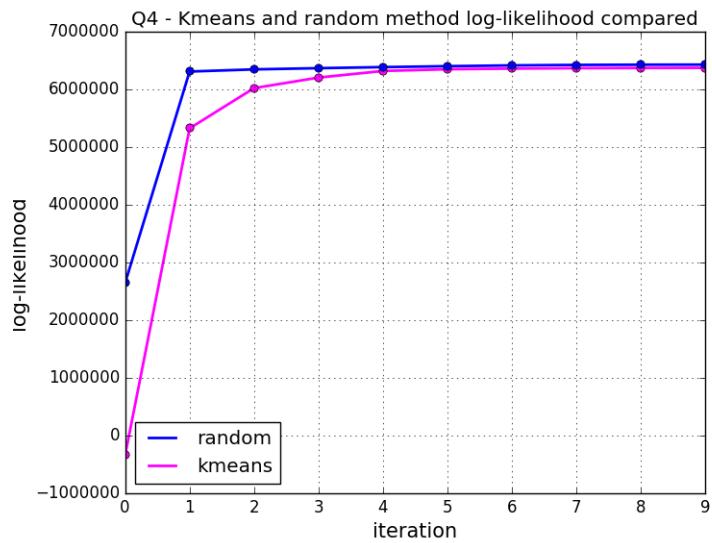


Figure 27: Comparison between log-likelihood using two different methods: Kmeans shown in magenta and randomized method (using randConst) shown in blue. We can see that using means method the log-likelihood converges earlier to a similar value compared to the randomized method.

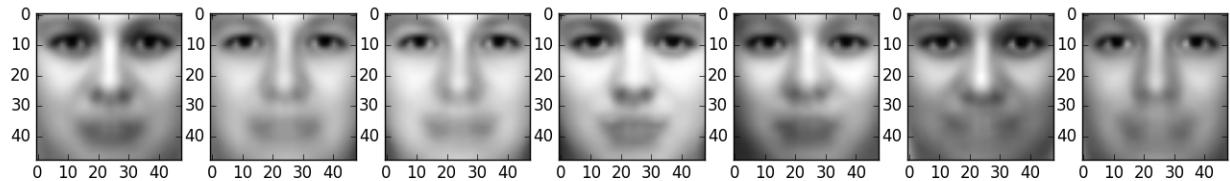


Figure 28:

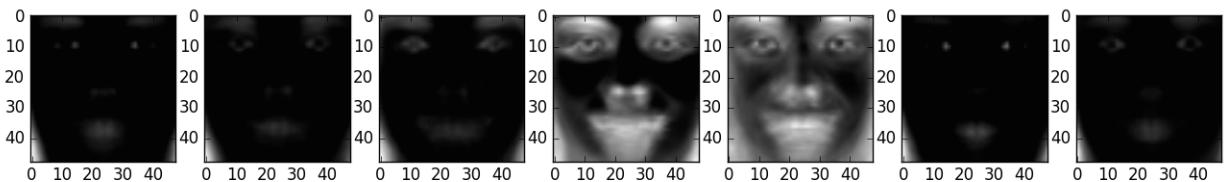


Figure 29:

4.4

In this question, we are trying to compute $p(d|x)$ value using Baye's rule. using Baye's rule we know that:

$$p(d|x) = \frac{p(x|d)p(d)}{p(x)}, \quad (1)$$

where $p(x|d)$, $p(d)$ is the prior and $p(x)$ is the evidence. Since $p(x)$ is constant for all the values we can ignore it here. Then, we can take log of both sides of Equation (1) and write it as:

$$\log(p(d|x)) = \log(p(x|d)p(d)) = \log(p(x|d)) + \log(p(d)). \quad (2)$$

We do have value of $p(d)$ from "log_likelihood_class" function in the code provided and value of $p(x|d)$ is given in the function called "mogLogLikelihood". So we can use these functions to compute $\log(p(d|x))$.

Figure ?? shows the for all three cases of training set, validation set and test set.

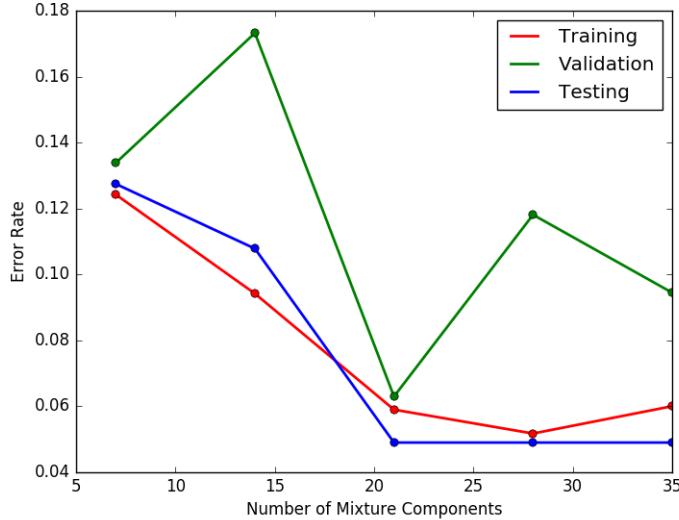


Figure 30:

Answers to the questions:

(b) We find that the error rates on the training set generally decreases as the number of clusters increases. This is because it is easier to fit to larger number of Gaussians. It obviously over-fits the training data if the number of cluster is very large but it still fits the training set well so the error rate decreases.

(c) The error rate for test set decreases as we increase the number of clusters. This is not the case for validation set as we see (the error rate increases for higher cluster numbers). This is because we are over-fitting the training set by having larger number of clusters and therefore the model does not work well on the validation set for larger number of clusters and the reason it works for the test set is because of the input we have for test set. I think we are just lucky that it works well on the test set and usually this is not the case.