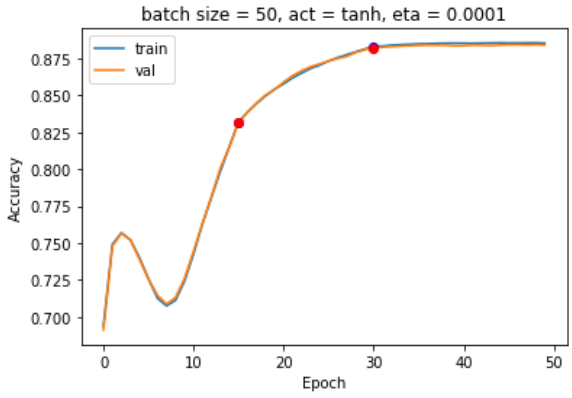
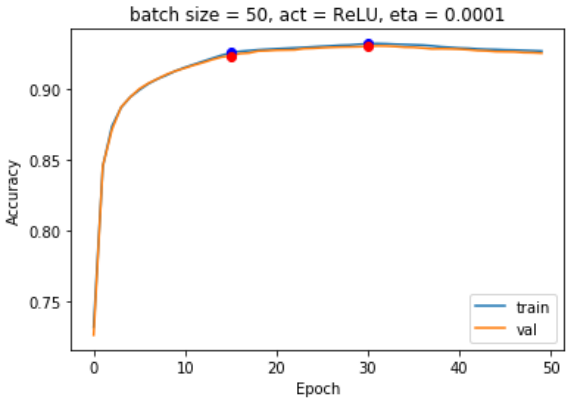
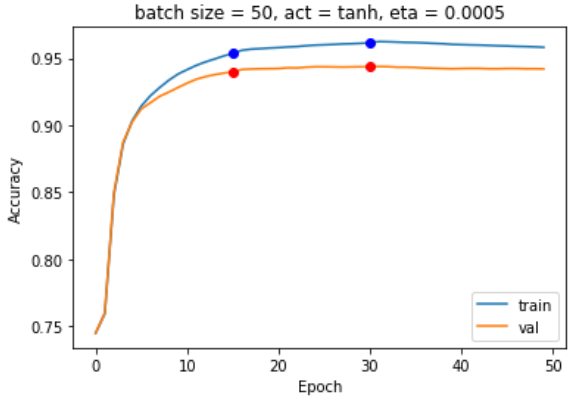
Problem 1



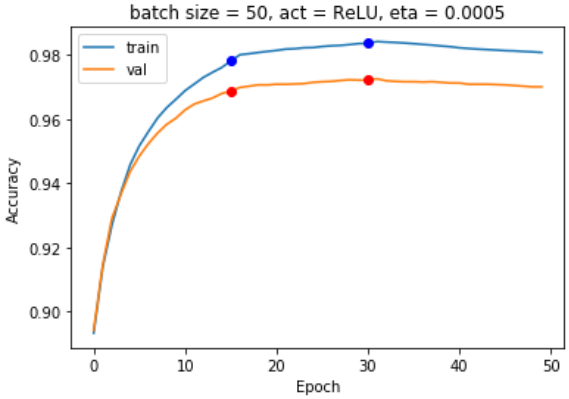
(a) activation: tanh, eta: 0.0001, accuracy on test set: 88.33%



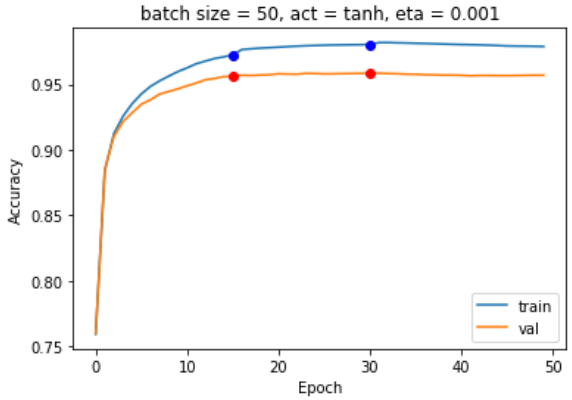
(d) activation: ReLU, eta: 0.0001, accuracy on test set: 92.98%



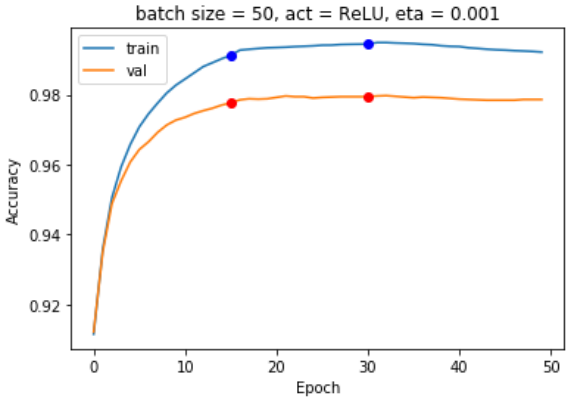
(b) activation: tanh, eta: 0.0005, accuracy on test set: 94.21%



(e) activation: ReLU, eta: 0.0005, accuracy on test set: 97.36%



(c) activation: tanh, eta: 0.001, accuracy on test set: 95.5%



(f) activation: ReLU, eta: 0.001, accuracy on test set: 97.87%

Fig. 1 Training Curves

Network configuration: There are four layers in my MLP, 784 neurons in the input layer, 400 neurons in the first hidden layer, 200 neurons in the second hidden layer and 10 neurons in the output layer.

Batch size: 50

Initial learn rate: 1e-5, 5e-4, 1e-4

Parameter initialization: weights is initialized with He initialization, biases are initialized as 0.

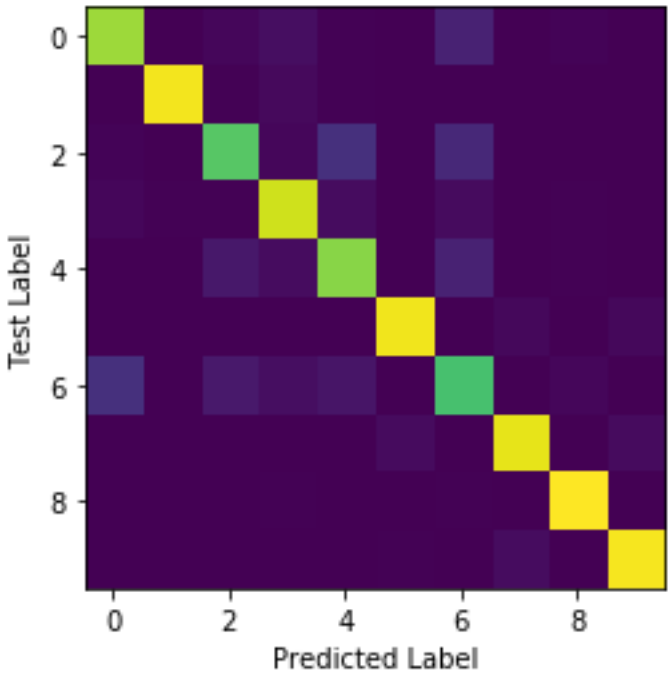
Best final accuracy: accuracy on test set is 97.95% while training with the whole training set, while training, accuracy on training set is 99.216%, accuracy on validation set is 97.86%, accuracy on test set is 97.87%

6 curves are shown in Fig. 1, the epoch to change learning rate is 15, 30 as shown in the plot with right and blue dots.

Problem 2

(1)

		Predicted Label									
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Test Label	**0**	858	0	12	18	4	0	97	1	10	0
	1	3	962	2	24	3	0	3	0	3	0
	2	17	0	841	7	84	0	48	0	3	0
	3	30	8	23	873	28	0	33	0	5	0
	4	0	0	153	30	762	0	54	0	1	0
	5	0	0	0	0	0	956	0	28	2	14
	6	148	1	116	18	78	0	625	0	14	0
	7	0	0	0	0	0	14	0	965	1	20
	8	3	0	4	3	2	1	6	6	975	0
	9	0	0	0	0	0	5	1	42	0	952



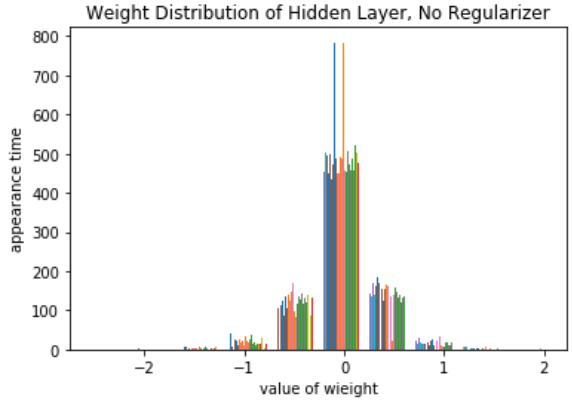
Confusion Matrix and Heat Map

(2)

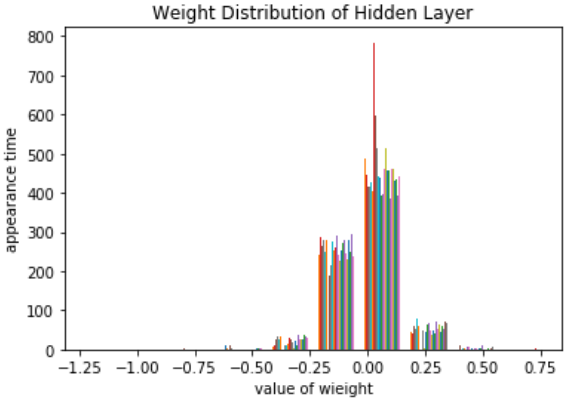
Test Label	0	1	2	3	4	5	6	7	8	9
Confused Label	6	3	4	6	2	7	0	9	6,7	7

(3) From the confusion matrix, the two classes are most likely to be confused is class 6 and 0.

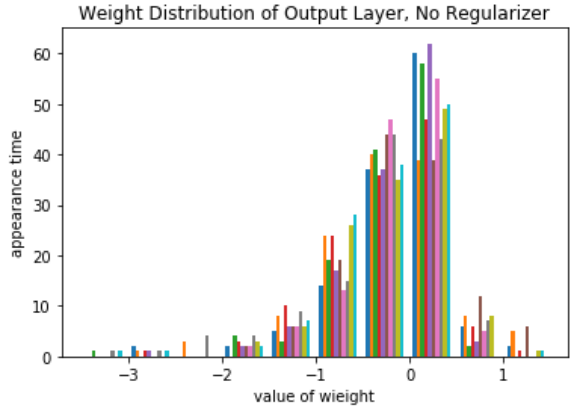
Problem 3



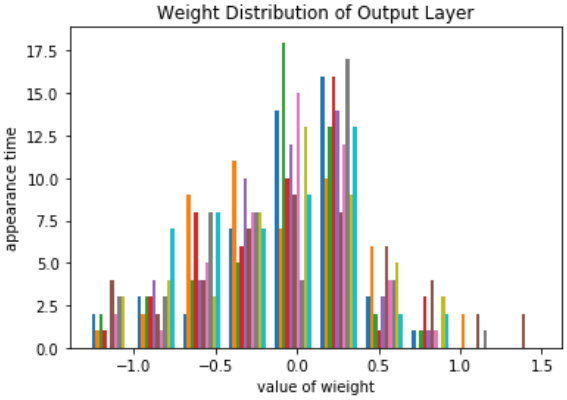
(a) Weight distribution of hidden layer (No Regularizer)



(c) Weight distribution of hidden layer (With Regularizer)



(b) Weight distribution of output layer (No Regularizer)

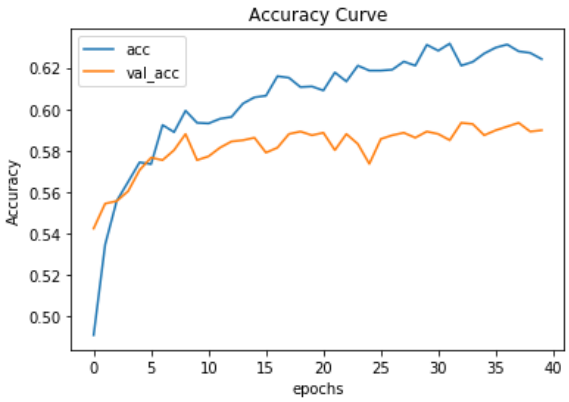


(d) Weight distribution of output layer (With Regularizer)

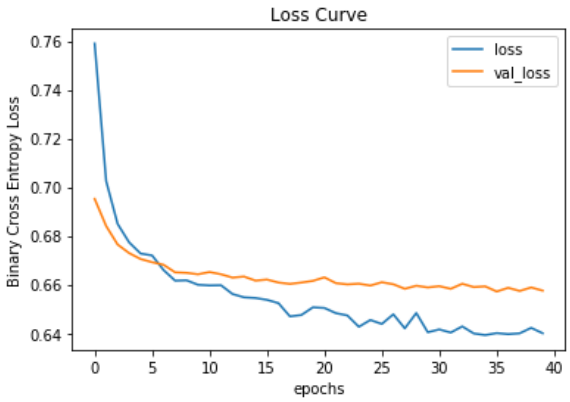
Fig. 2 Weight Distribution

From the histogram above, the MLP without regularizer and more neurons each layer tends to have more weights with absolute value larger than 1 in the hidden layer, also more weights with absolute value larger than 1 in the output layer, while the MLP with regularizer tend to have weights with smaller absolute value. Besides, the MLP without regularizer and more neurons tend to have a more symmetrical distribution in the hidden layer than that of the MLP with regularizer and less neurons.

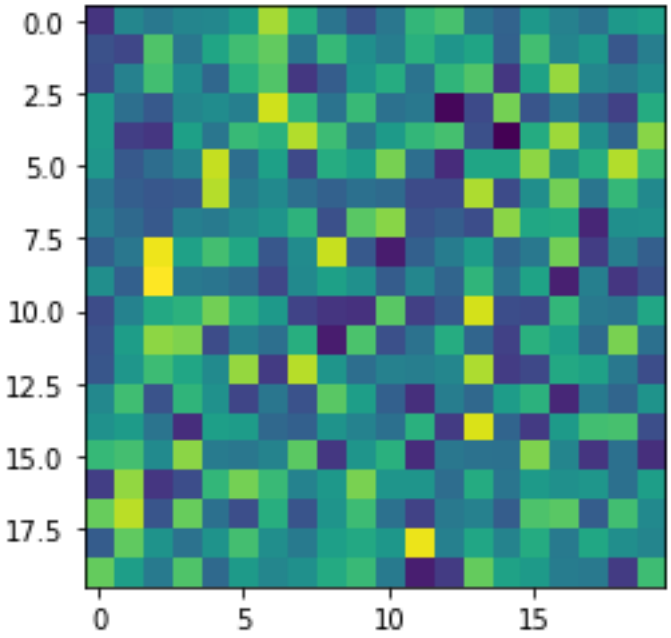
Problem 4



(a) Accuracy Curve



(b) Loss Curve



(c) Weight Matrix of hidden layer

Fig. 3 Problem 4

From the weight matrix of hidden layer, there may be some significant information in the 3rd, 4th, 8th, 11th, 13th feature, where there are some larger values of weight.