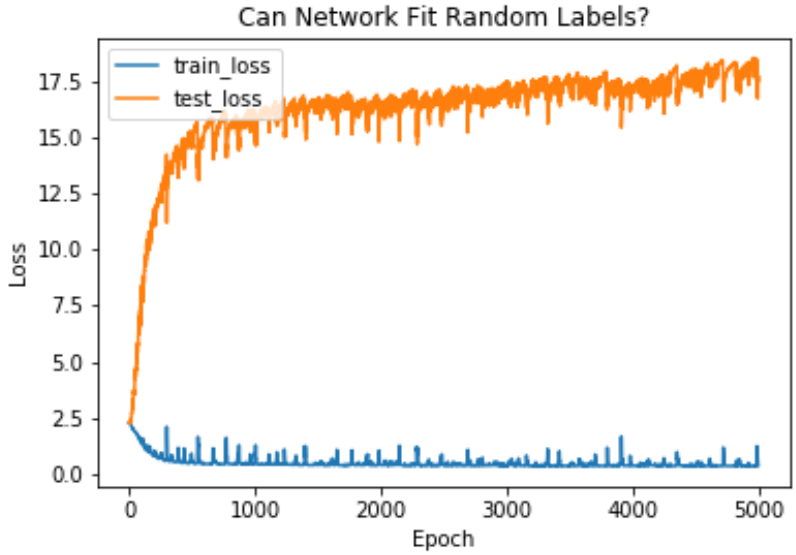
HW 1-3 Generalization

1. Can network fit random variables?

- **Describe the experiment settings:**

  I train the classification task on MNIST dataset with a simple DNN model, which epoch is 5000. The training data size is 1000 and batch size is 100. It can let the model fit the training data (Overfitting) easily. The number of parameters is 16750, and the learning rate is 3e-3.

- **Plot the figure of the relationship between training and testing, loss and epochs:**

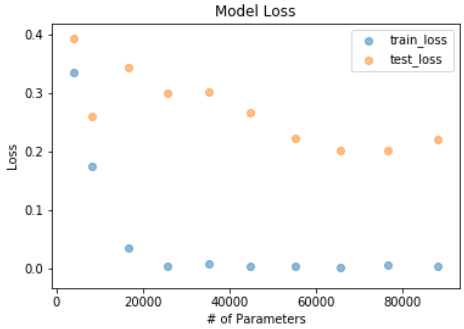


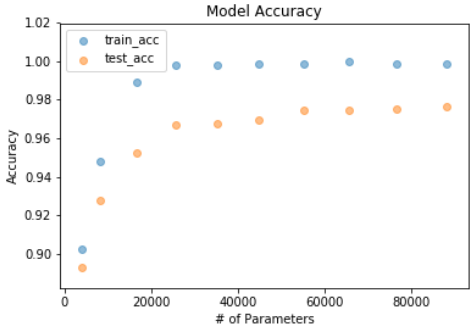
2. Number of parameters v.s. Generalization

- **Describe the experiment settings:**

  I chose the classification task on MNIST dataset with all training data (55000 images). The learning rate, epoch, batch size is 3e-3, 100 and 200, respectively. Besides, I trained on 10 different model structures. They both have 3 hidden layers. The neurons in hidden layers of each structures are 5, 10, 20, 30, 40, 50, 60, 70, 80 and 90.

- **Plot the figures of both training and testing, loss and accuracy to the number of parameters:**





- **Comment the result:**

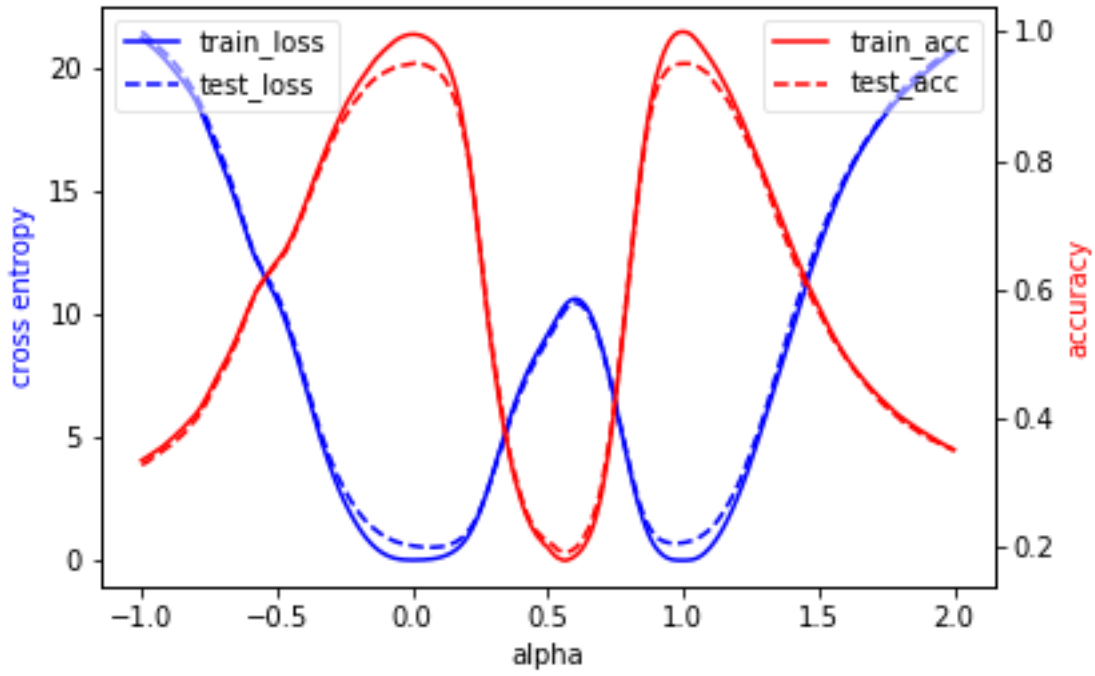
  I found that even when the number of parameters is enough for training dataset, the generalization ability can be better when the parameters getting more. In my concept, the model may be over fitting when the number of parameters is too large, which may lead to bad testing accuracy. However, the result of this experiment shows the opposite result.

3. Flatness v.s. Generalization – part1

- **Describe the experiment settings:**

  I chose the classification task on MNIST dataset with all training data (55000 images). The two different model structures include different batch size and learning rate with same epoch which is 500. The different batch size and learning rate are 1e-3, 1e-2 and 64, 1024 respectively.

- **Plot the figures of both training and testing, loss and accuracy to the number of interpolation ratio:**



- **Comment the result:**

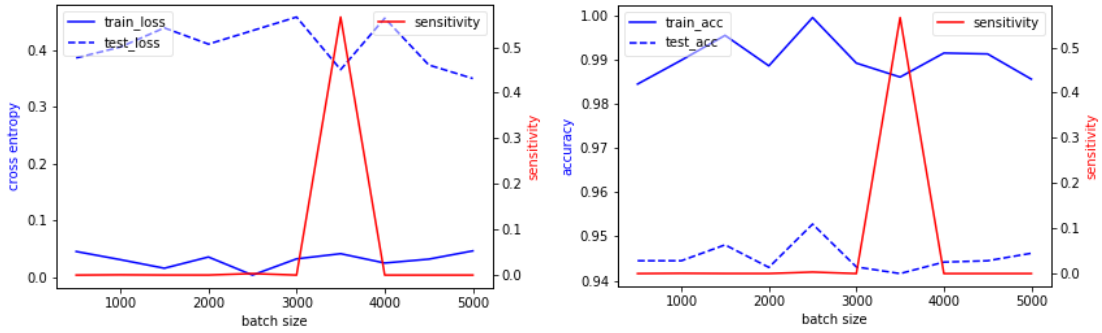
  As expected, the solution of two different model structures are different local minima. The interpolation of two weights show the peak through loss surface. The performance is not good when alpha is 0.5 (i.e. the middle point between two different local minima).

4. Flatness v.s. Generalization – part2

- **Describe the experiment settings:**

I chose the classification task on MNIST dataset with all training data (55000 images). The learning rate and epoch are 1e-2 and 200 respectively. I trained on ten different model structures with batch size 500, 1000, 1500, 2000 … 5000. The sensitivity is defined as 2-norm Jacobian Matrix with a random input x (i.e. 784-dimension vector).

- **Plot the figures of both training and testing, loss and accuracy, sensitivity to your chosen variable:**



- **Comment the result:**

The sensitivity getting larger when the training/testing accuracy getting smaller. When the batch size is 3500, the sensitivity is very high and the accuracy is small. Though the training loss is high too, the testing loss somehow decrease. This makes sense excludes the testing loss. In my concept, the testing loss would bet large when the training loss is large too.