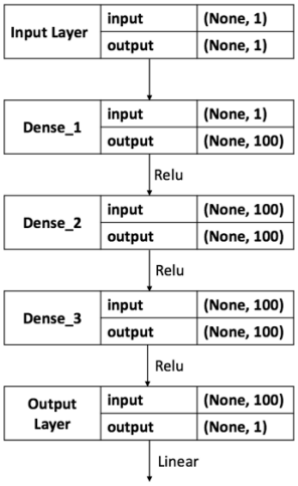
HW 1-1 Deep v.s. Shallow

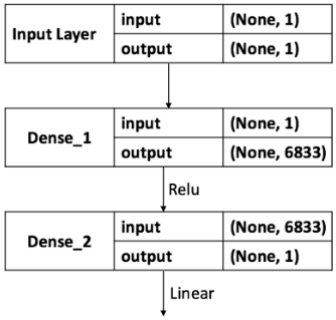
1. Simulate a function

- **Describe the model and the function you used:**

  I use two models with different layers, and the parameters are both around 20500. Each activation function of the hidden layer is Relu, and the activation function of output layer is linear.



Deep Model
(parameters: 20501)

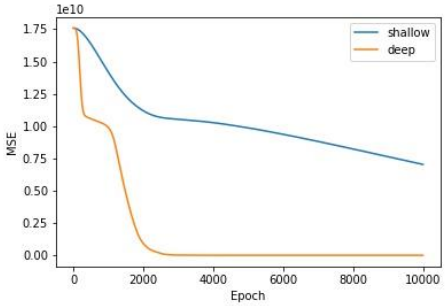


Shallow Model
(parameters: 20500)

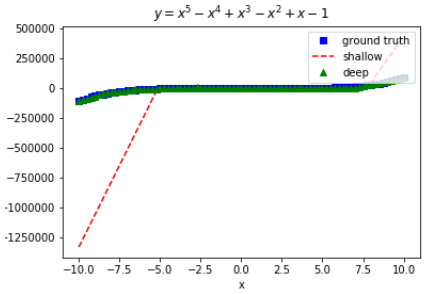
$$y = x^5 - x^4 + x^3 - x^2 + x - 1$$

Objective Function

- **Plot the training loss of all models:**



- **Plot the predicted function curve of all models and ground truth:**



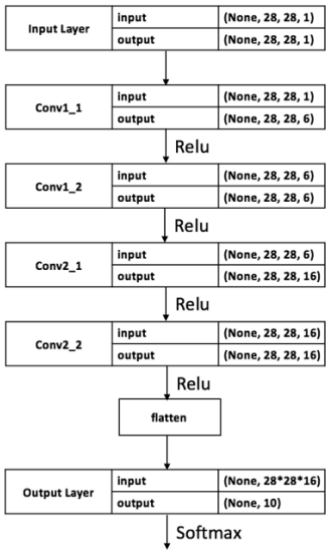
- **Comment the result:**

  In this task, I found that there are still large of basic knowledges about DL that I missed.

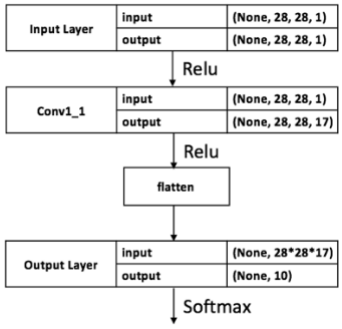
Frist, the initialization of each kernel is very important since it may lead to different critical point. When the model gets deeper and more complicated, the prediction may lead to flat loss curve and stuck in the saddle point if the initializations are near from the original point. Second, this is the regression problem and the activation function of the output layer cannot be Relu!! At the first time of this task, I cannot even train the ability of this model and I finally found that I use Relu activation function of the output layer when I try to visualize the prediction function. It really looks like a… Relu function, which is far from the objective function.

2. Train on actual task

- **Describe the model and the task you choose:**
  The task I chose is for classification on MNIST.

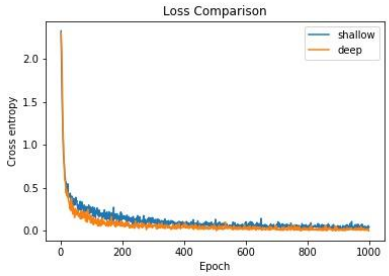


Deep Model
(parameters: 135344)

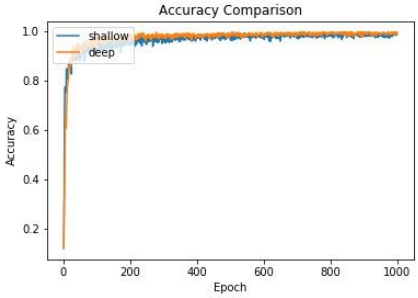


Shallow Model
(parameters:133732)

- Plot the training loss of all models:



- Plot the training accuracy of all models:



- Comment the result:
  In this task, the result seems normal. When the model gets deeper, the loss value can

decrease faster, and the accuracy can increase faster too. However, the final performance of deep and shallow model seem similar. I think that it is because the dataset I chose is MNIST, which is too simple for such the shallow model. If the dataset getting more complicate like CIFAR-10, the difference between two models may get larger.