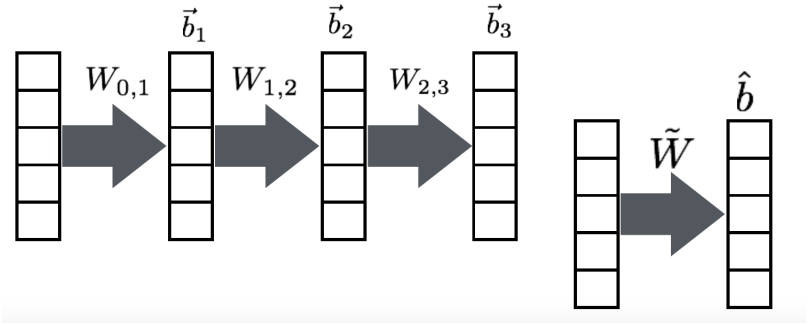
CS5242: NEURAL NETWORKS AND DEEP LEARNING

Assignment 1

Question 1



For the left NN, assuming all activation functions are identity functions,

$$\vec{h_1} = \sigma(W_{0,1} \cdot \vec{x} + \vec{b_1}) = W_{0,1} \cdot \vec{x} + \vec{b_1} \tag{1}$$

$$\vec{h_2} = \sigma(W_{1,2} \cdot \vec{h_1} + \vec{b_2}) = W_{1,2} \cdot \vec{h_1} + \vec{b_2} \tag{2}$$

$$\vec{o} = \vec{h_3} = \sigma(W_{2,3} \cdot \vec{h_2} + \vec{b_3}) = W_{2,3} \cdot \vec{h_2} + \vec{b_3} \tag{3}$$

From equations (1), (2) and (3), we get

$$\begin{aligned}
\vec{o} &= W_{2,3} \cdot \vec{h_2} + \vec{b_3} \\
&= W_{2,3} \cdot (W_{1,2} \cdot \vec{h_1} + \vec{b_2}) + \vec{b_3} \\
&= W_{2,3} \cdot (W_{1,2} \cdot (W_{0,1} \cdot \vec{x} + \vec{b_1}) + \vec{b_2}) + \vec{b_3} \\
&= W_{2,3} \cdot (W_{1,2} \cdot W_{0,1} \cdot \vec{x} + W_{1,2} \cdot \vec{b_1} + \vec{b_2}) + \vec{b_3} \\
&= W_{2,3} \cdot W_{1,2} \cdot W_{0,1} \cdot \vec{x} + W_{2,3} \cdot W_{1,2} \cdot \vec{b_1} + W_{2,3} \cdot \vec{b_2} + \vec{b_3}
\end{aligned} \tag{4}$$

For the right NN without any hidden layer, we have

$$\vec{o} = \bar{W} \cdot \vec{x} + \hat{b} \tag{5}$$

Assuming the two neural networks are equivalent, from (4) and (5), we get

$$\hat{b} = W_{2,3} \cdot W_{1,2} \cdot \vec{b_1} + W_{2,3} \cdot \vec{b_2} + \vec{b_3} \tag{6}$$

$$\bar{W} = W_{2,3} \cdot W_{1,2} \cdot W_{0,1} \tag{7}$$

Question 2

The three neural nets are trained using a Mini-Batch Gradient Descent algorithm using a batch size of 25. To keep things simple, regularizing (λ) and learning rate annealing are not introduced to the nets. In this setup, it is found that 14-14x28-4 net has the best performance and is the most suitable for the task at hand as explained below.

The first neural net, 14-100-40-4 net is a wide, shallow neural net which converges earlier than the others while providing very good cost and accuracy scores. However, it has also overfitted the data as evidenced by Figure 3. This can be prevented by adding a regularization parameter to the weights during the gradient descent.

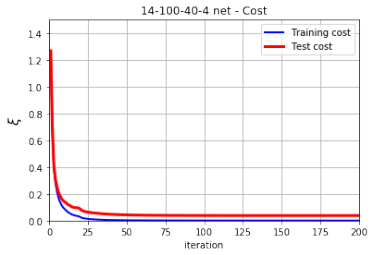


Fig.1 (Net 1 Cost)

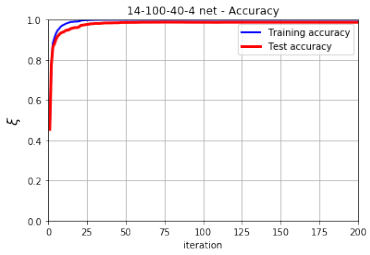


Fig.2 (Net 1 Accuracy)

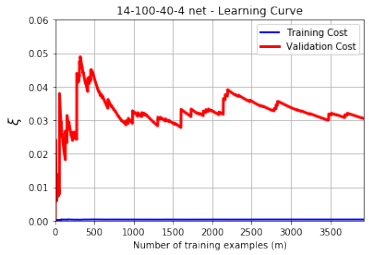


Fig.3 (Net 1 Learning Curve)

The second neural net, 14-28x6-4 net is a 6-layers deep neural net. Similar to the first net, this net also suffers from a high variance issue. The performance can be improved by regularizing or adding more training examples.

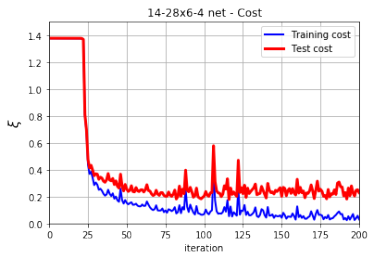


Fig.4 (Net 2 Cost)

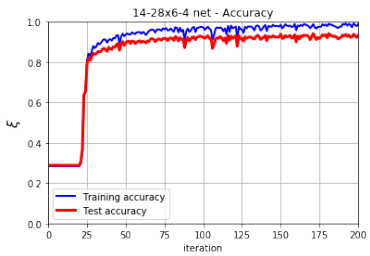


Fig.5 (Net 2 Accuracy)

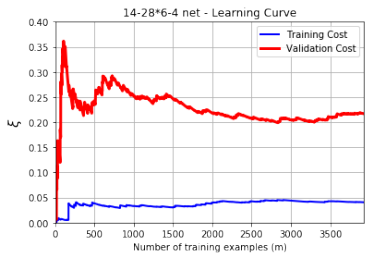


Fig.6 (Net 2 Learning Curve)

The third neural net, 14-14x28-4 net is a 28-layers deep neural net. The accuracy of this net is around 92% and as shown in figure 9, this net neither overfits nor underfits the data. By having more layers, it can abstract the data better than the shallower networks. The error spikes in Figure 7 can be reduced by annealing the learning rate.

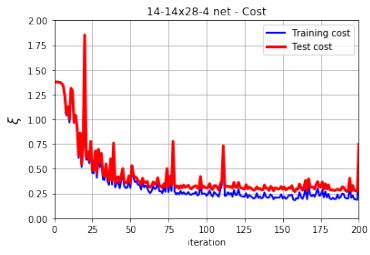


Fig.7 (Net 3 Cost)

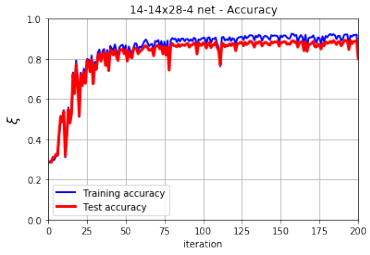


Fig.8 (Net 3 Accuracy)

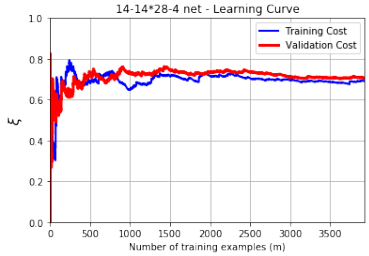


Fig.9 (Net 3 Learning Curve)