Deep Learning HW2

Question 1 - Generalization in A Teacher-Student Setup

$$x^{(0)} \sim N(0, I)$$

$$R(w) = E_{x(0)}[\left|\left|w^T x^{(0)} - w_t^T x^{(0)}\right|\right|^2]$$

Prove:

$$R(w) = \left|\left|w - w_t\right|\right|^2$$

Solution:

$$R(w) = E_{x(0)}[\left|\left|w^T x^{(0)} - w_t^T x^{(0)}\right|\right|^2]$$

$$= E_{x(0)}\left[\sum_{i=1}^{d}\left(w_i x_i - w_{i,t} x_i\right) * \sum_{j=1}^{d}\left(w_j x_j - w_{j,t} x_j\right)\right]$$

$$= E_{x(0)}\left[\sum_{i=1}^{d}\left(w_i x_i - w_{i,t} x_i\right)^2 + \sum_{j=1}^{d}\left(w_j x_j - w_{j,t} x_j\right) * \sum_{i!=j}^{d}\left(w_i x_i - w_{i,t} x_i\right)\right]$$

$$= \sum_{i=1}^{d}\left(w_i - w_{i,t}\right)^2 E_{x(0)}\left[x_i^2\right] + \sum_{j=0}^{d}\sum_{i!=j}^{d}\left(w_j - w_{j,t}\right)\left(w_i - w_{i,t}\right) E_{x(0)}[x_i x_i]$$

$$*** \; E_{x(0)}[x_i x_i] = Cov\left(x_i, x_j\right) = 0 \rightarrow Given\ fact$$

$$= Var(x_i)\sum_{i=1}^{d}\left(w_i - w_{i,t}\right)^2 = I||w - w_t||^2$$

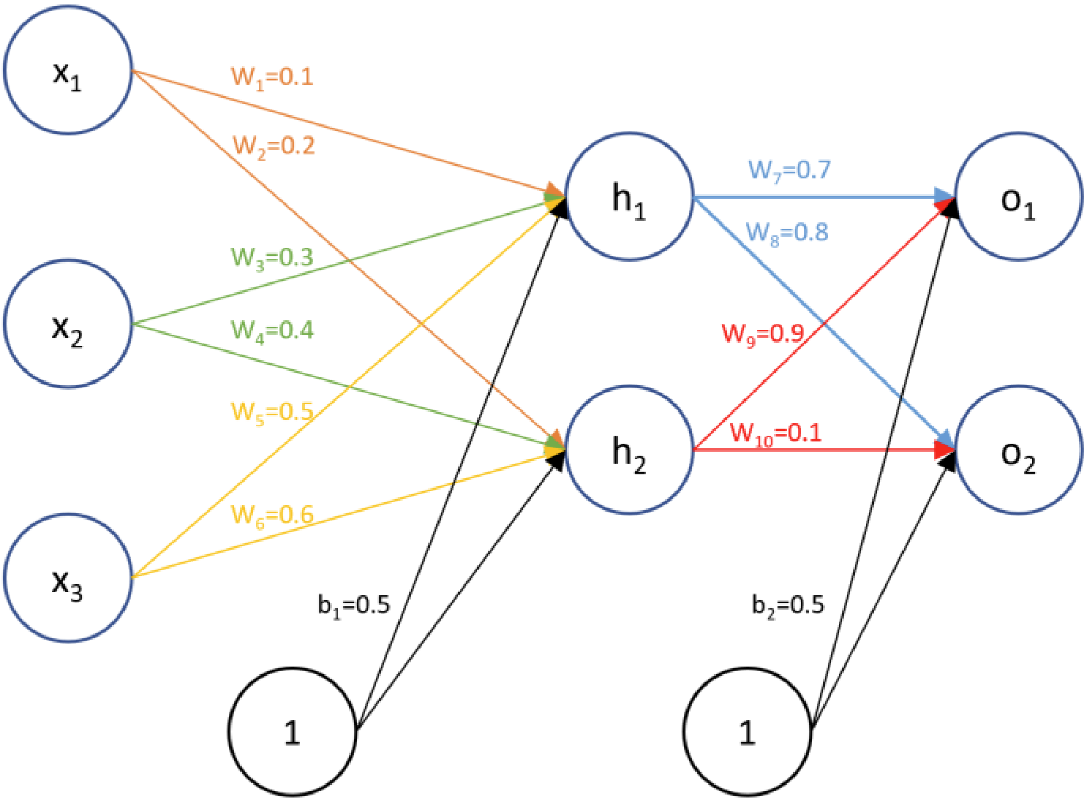
Question 2 – Backpropagation By Hand

$$x_1, x_2, x_3, = [1,4,5]$$

$$t = [t_1, t_2] = [0.1, 0.05]$$

$$Activation = Sigmoid$$

$$Loss = MSE\ (Mean\ Squared\ Error)$$



1. **Perform forward pass, and calculate the MSE**

$$h_1 = sigmoid(x_1 w_1 + x_2 w_3 + x_3 w_5 + b_1) = 0.986613$$

$$h_2 = sigmoid(x_1 w_2 + x_2 w_4 + x_3 w_6 + b_1) = 0.995033$$

$$o_1 = w_7 h_1 + w_9 h_2 + b_2 = 0.889550$$

$$o_2 = w_8 h_1 + w_{10} h_2 + b_2 = 0.800399$$

$$MSE = \frac{1}{2}(o_1 - t_1)^2 + \frac{1}{2}(o_2 - t_2)^2 = 0.593244$$

2. **Perform backward pass, and calculate the gradients**

$$\frac{d\,MSE}{do_1} = \frac{1}{2} * 2 * (o_1 - t_1) = 0.79$$

$$\frac{d\,MSE}{do_2} = \frac{1}{2} * 2 * (o_2 - t_2) = 0.75$$

$$\frac{d\,MSE}{dw_9} = \frac{d\,MSE}{do_1} * \frac{do_1}{dw_9} = \frac{1}{2} * 2 * (o_1 - t_1) * h_2 * o_1 * (1 - o_1) = 0.074$$

$$\frac{d\,MSE}{dw_{10}} = \frac{d\,MSE}{do_2} * \frac{do_2}{dw_{10}} = \frac{1}{2} * 2 * (o_2 - t_2) * h_2 * o_2 * (1 - o_2) = 0.117$$

$$\frac{d\,MSE}{dw_7} = \frac{d\,MSE}{do_1} * \frac{do_1}{dw_7} = \frac{1}{2} * 2 * (o_1 - t_1) * h_1 * o_1 * (1 - o_1) = 0.0765$$

$$\frac{d\,MSE}{dw_8} = \frac{d\,MSE}{do_2} * \frac{do_2}{dw_8} = \frac{1}{2} * 2 * (o_2 - t_2) * h_1 * o_2 * (1 - o_2) = 0.118$$

$$\frac{do_1}{dh_1} = w_7 * o_1(1 - o_1) = 0.068$$

$$\frac{do_1}{dh_2} = w_9 * o_1(1 - o_1) = 0.088$$

$$\frac{do_2}{dh_1} = w_8 * o_2(1 - o_2) = 0.128$$

$$\frac{do_2}{dh_2} = w_{10} * o_2(1 - o_2) = 0.016$$

$$\frac{d\,MSE}{dw_1} = \frac{d\,MSE}{do_1} * \frac{do_1}{dh_1} * \frac{dh_1}{dw_1} + \frac{d\,MSE}{do_1} * \frac{do_1}{dh_1} * \frac{dh_1}{dw_1} = 0.002$$

$$\frac{d\,MSE}{dw_3} = \frac{d\,MSE}{do_2} * \frac{do_2}{dh_1} * \frac{dh_1}{dw_3} + \frac{d\,MSE}{do_1} * \frac{do_1}{dh_1} * \frac{dh_1}{dw_3} = 0.008$$

$$\frac{d\,MSE}{dw_5} = \frac{d\,MSE}{do_2} * \frac{do_2}{dh_1} * \frac{dh_1}{dw_5} + \frac{d\,MSE}{do_1} * \frac{do_1}{dh_1} * \frac{dh_1}{dw_5} = 0.099$$

$$\frac{d\,MSE}{dw_2} = \frac{d\,MSE}{do_2} * \frac{do_2}{dh_2} * \frac{dh_2}{dw_2} + \frac{d\,MSE}{do_1} * \frac{do_1}{dh_2} * \frac{dh_2}{dw_2} = 0.00095$$

$$\frac{d\,MSE}{dw_4} = \frac{d\,MSE}{do_2} * \frac{do_2}{dh_2} * \frac{dh_2}{dw_4} + \frac{d\,MSE}{do_1} * \frac{do_1}{dh_2} * \frac{dh_2}{dw_4} = 0.0037$$

$$\frac{d\,MSE}{dw_6} = \frac{d\,MSE}{do_2} * \frac{do_2}{dh_2} * \frac{dh_2}{dw_6} + \frac{d\,MSE}{do_1} * \frac{do_1}{dh_2} * \frac{dh_2}{dw_6} = 0.00474$$

$$\frac{d\,MSE}{db_2} = \frac{d\,MSE}{do_2} * \frac{do_2}{db_2} = 0.12$$

$$\frac{d\,MSE}{db_1} = \frac{d\,MSE}{do_2} * \frac{do_2}{dh_2} * \frac{dh_2}{db_1} + \frac{d\,MSE}{do_1} * \frac{do_1}{dh_2} * \frac{dh_2}{db_1} + \frac{d\,MSE}{do_2} * \frac{do_2}{dh_1} * \frac{dh_1}{db_1} + \frac{d\,MSE}{do_1} * \frac{do_1}{dh_1} * \frac{dh_1}{db_1} = 0.00292$$

3. Calculate the new weights with SGD with learning rate of 0.01

Foreach parameter x we apply the following equation: $x_{t+1} = x_t - a\frac{d\,MSE}{dx}$ and get:

$$w_1^{t+1} = w_1 - a\frac{d\,MSE}{dw_1} = 0.0998$$

$$w_2^{t+1} = w_2 - a\frac{d\,MSE}{dw_2} = 0.1999$$

$$w_3^{t+1} = w_3 - a\frac{d\,MSE}{dw_{31}} = 0.2999$$

$$w_4^{t+1} = w_4 - a\frac{d\,MSE}{dw_4} = 0.3999$$

$$w_5^{t+1} = w_5 - a\frac{d\,MSE}{dw_5} = 0.4999$$

$$w_6^{t+1} = w_6 - a\frac{d\,MSE}{dw_6} = 0.5999$$

$$w_7^{t+1} = w_7 - a\frac{d\,MSE}{dw_7} = 0.6923$$

$$w_8^{t+1} = w_8 - a\frac{d\,MSE}{dw_8} = 0.7988$$

$$w_9^{t+1} = w_9 - a\frac{d\,MSE}{dw_9} = 0.8992$$

$$w_{10}^{t+1} = w_{10} - a\frac{d\,MSE}{dw_{10}} = 0.0988$$

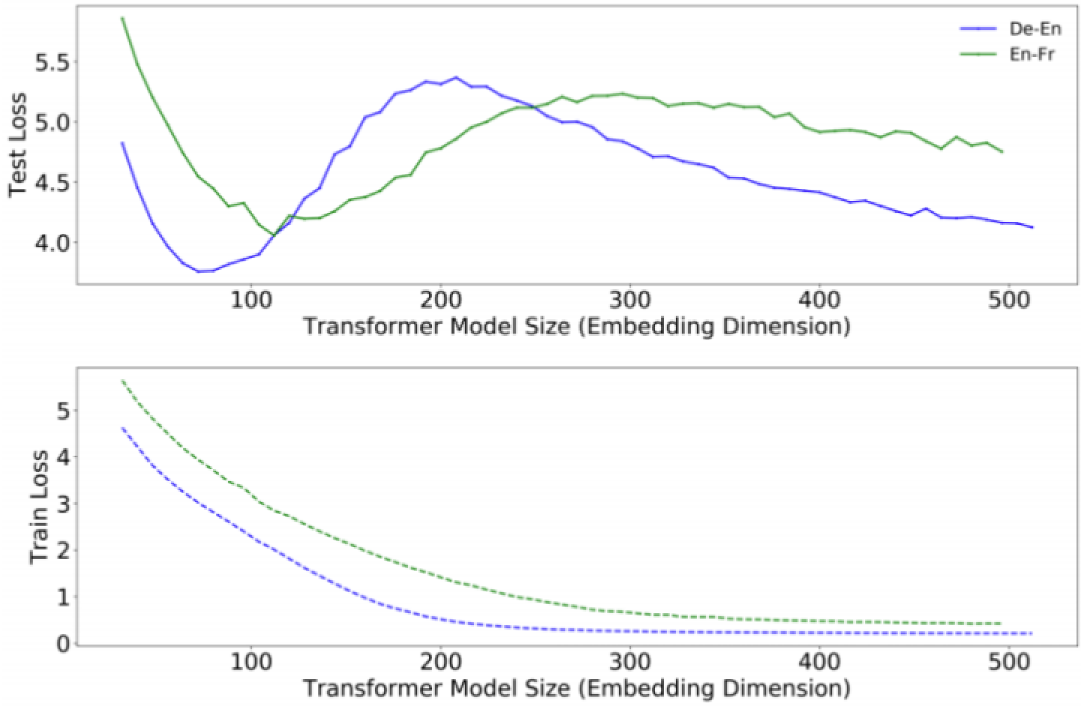
$$b_1^{t+1} = b_1 - a\frac{d\,MSE}{db_1} = 0.4999$$

$$b_2^{t+1} = b_2 - a\frac{d\,MSE}{db_2} = 0.4988$$

Question 3 – Deep Double Decent

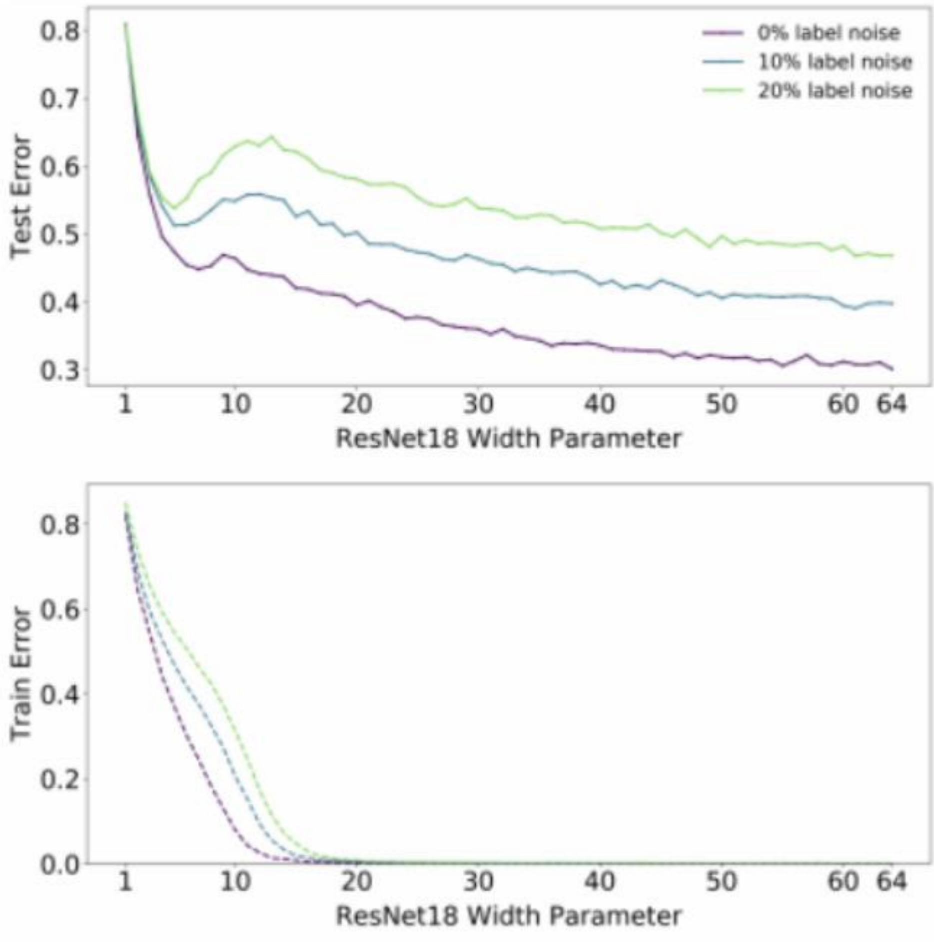
1. **Where is the critical point?**
2. **What type of double descent is shown?**

a. **Part A**



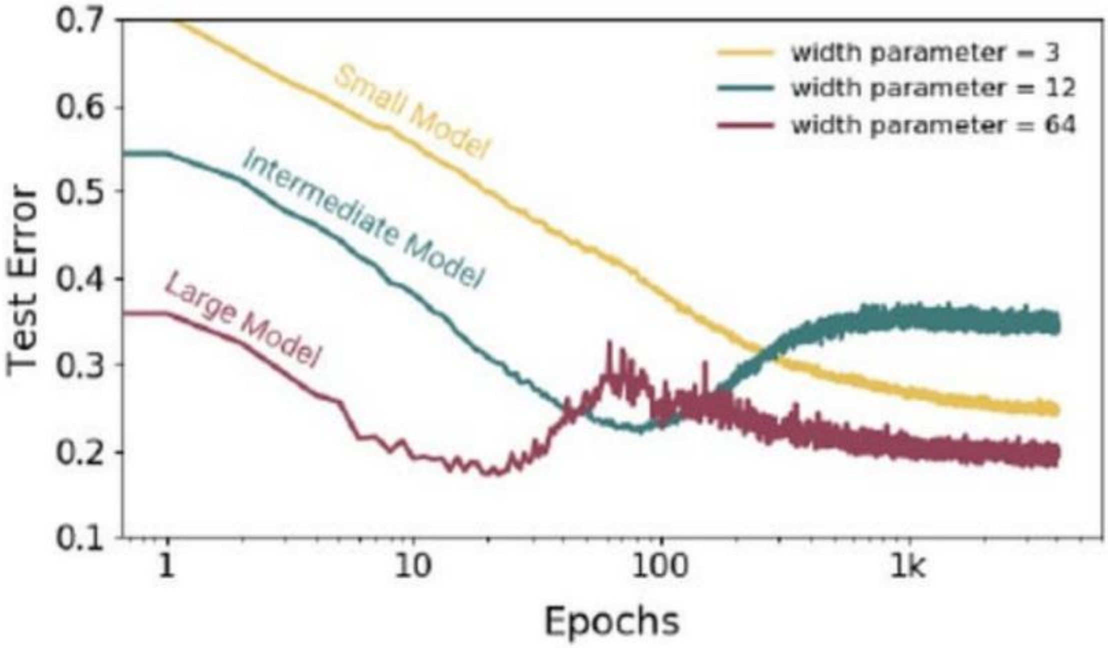
1. The critical point is when the first plot reaches 200, and for the second graph when reaching 300.
2. Model double decent.

b. Part B



1. The critical point is located approximately between 10 to 14 for both graphs.
2. Model double decent

c. **Part C**



1. Small models don't suffer from this phenomenon, hence no critical point in the yellow model, same goes for the intermediate model where it reaches a stable maximum at higher epochs, however larger models suffer quite a bit, the red has a critical point in between 70 and 90 epochs.
2. Epoch-wise double descent. Reverse overfitting on the intermediate model.

Question 4 – Initialization

From the lecture we concluded that when for large widths $z \sim N(01)$ and the following holds

$$E_z[\varphi^2(z)] = E_z[\max(z, 0)^2]$$

$$P_u\ is\ symetric\ hence\ P_z\ is\ also\ symetric$$

Hence

$$\int_{-\infty}^{\infty} \max(z, 0)^2 f_Z(z) = P(z > 0) \int_0^{\infty} \max(z, 0)^2 f_Z(z)dz + P(z < 0) * 0 = \frac{1}{2}\int_0^{\infty} z^2 f_Z(z)dz$$
$$= \frac{1}{2}E[z^2] = \frac{1}{2}Var(z) = \frac{1}{2}$$

The width of the previous layer is d_{l-1}we get the following:

$$\sigma_l = \frac{1}{\sqrt{\sum_i E[\varphi^2(u_{l-1}[i])}} \cong \frac{1}{\sqrt{d_{l-1}E_z[\varphi^2(z)]}} = \frac{1}{\sqrt{d_{l-1} * 0.5}} = \sqrt{\frac{2}{d_{l-1}}}$$

Question 5 – MLP and Invariance

1. The activation function (Leaky RELU) is the only invariance in this network when $0 < \rho < 1$

$$LeakyRele(cx) = \max(pcx, cx) *= c * \max(px, x) = c * LeakyRelu$$

*** the *= is possible since the arguments of the max function are two linear functions.

2. For this case we need to force the parameters for each row in the input matrix to be symmetric meaning the first and last element parameter should be the same along the d-axis.
   Foreach row in input matrix:
   Foreach col in range(d/2) :
   w[row,col] == w[row, d-col]

3. We need to force the same parameters across equivalent rows, meaning all parameters in row 1 are the same as the parameters in row 4, same goes for rows 2 and 3.

4. (a) The network is already trained so the positive neurons will still fire a positive result and the negative neurons will still fire a negative result, since the network is only a single layer no propagated change takes effect, hence no change at all in the classification.

   (b) the learning rate is multiplied by c hence we expect the algorithm to never converge if the multiplication c*a is greater than 1.

   (c) one way to eliminate the effect of this constant is by making each column a one hot encoding only one row is turned on.

Question 6 – VGG Architecture

1. Complete the network architecture

Layer	Output Dimension	Number of parameters
INPUT	224x224x3	0
CONV3-64	224x224x64	1792
RELU	224x224x64	0
POOL2	112x112x64	0
CONV3-128	112x112x128	73856
RELU	112x112x128	0
POOL2	56x56x128	0
CONV3-256	56x56x256	295168
RELU	56x56x256	0
CONV3-256	56x56x256	590080
RELU	56x56x256	0
POOL2	28x28x256	0
CONV3-512	28x28x512	1180160
RELU	28x28x512	0
CONV3-512	28x28x512	2359808
RELU	28x28x512	0
POOL2	14x14x512	0
CONV3-512	14x14x512	2359808
RELU	14x14x512	0
CONV3-512	14x14x512	2359808
RELU	14x14x512	0
POOL	7x7x512	0
FC-4096	4096	102764544
FC-4096	4096	16781312
FC-1000	1000	4097000
SOFTMAX	1	0

2. Total number of params

For conv layers the number of parameters is: $3x3xIFxOF + OF$

For FC layers the number of parameters is: $IFxOF + OF$

Sum of parameters = 132,863,336

3. Percentage of fully connected layer params from overall params

$$ratio = \frac{123,642,856}{132,863,336} = 0.9306 \rightarrow 93.06\%$$