University of Science and Technology
Communications & Information Engineering Program
CIE 555: Neural Networks and Deep Learning
Spring 2020

مدينة زويل للعلوم والتكنولوجيا
Zewail City of Science and Technology

Progress Exam # 1 (60 min, 10 pts)

Name: ______________________ **ID:** ______________ **Grade:** ______

1) (1 points) Explain why it is preferred to use the cross-entropy cost function instead of the mean squared error when training neural networks to learn a probability distribution $p(y|x)$.

You may answer the question in different ways. A good answer shall include:

- **A good cost function shall produce a gradient that is large and predictable. Functions that saturate (gradient becomes very small in certain output regions) will not guide the training in an effective manner**
- **Many output units (e.g. sigmoid, softmax) involve an exp function**
- **The cross-entropy undoes the effect of the 'exp' function in these units**

2) (1 point) Explain the advantages/disadvantage of using Rectified Linear Units (ReLU) in the design of hidden units in deep feedforward networks.

A good answer may include the following advantages:

- **Easier to optimize (simpler and more efficient computationally).**
- **Gradients are large (when the unit is active) and consistent**
- **Sparsity: inactive nodes when the the weighted sum of input is less than zero.**

Disadvantages:

- **May result in an early saturation of nodes preventing them from updating their weight (vanishing gradient problem)**

3) (2 point) Discuss three generalizations of the ReLU units. Explain, using figures, why these generalized units might show better performance when training deep networks.

Examples: Leaky ReLU, Absolute value rectification, and Parametric ReLU. Please refer to the lecture for the discussion of these units.

These generalizations help with the vanishing gradient problems (when nodes become saturated early in the training process) by allowing units to have even if the weighted sum of unit's input is in the negative zones.

4) (2 points) Briefly discuss the difference between overfitting and underfitting and how they are related to the bias and variance in the model.

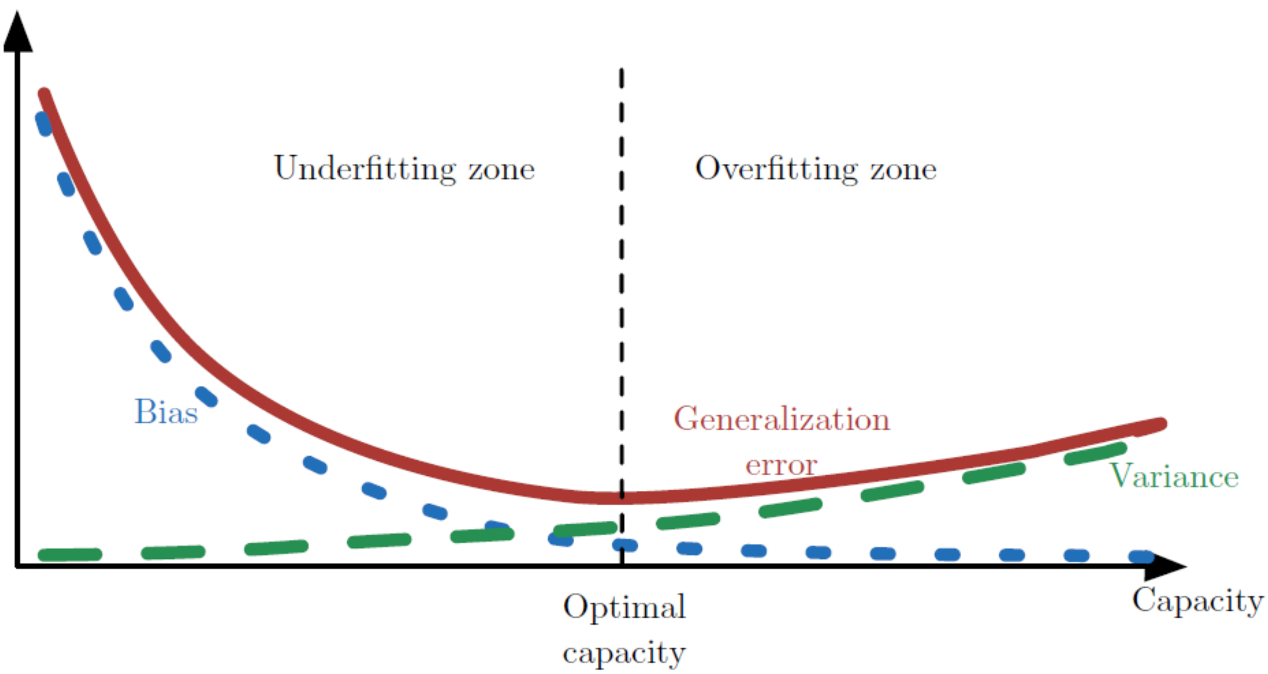
Overfitting: occurs when the gap between the training and test error increases

Underfitting: occurs when the model is not able to achieve a good accuracy on the training set.

Underfitting is typically associated with models of small capacity and high bias

Overfitting is typically associated with models of high capacity and high variance

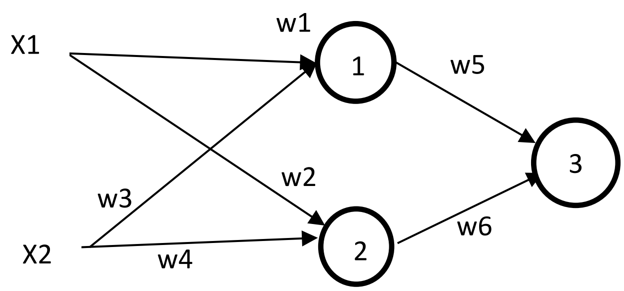
Progress Exam # 1 (60 min, 10 pts)



Deep Learning: Goodfellow, Bengio, Courville 2016

Progress Exam # 1 (60 min, 10 pts)

5) [4 points] Consider the following neural network



Give an explicit expression for the new (updated) weights w1, w2, w3, w4, w5 and w6 after backward propagation. Assume that the activation function used in all units is tanh.

Hint: Derivative of tanh(x) = 1 – $\tanh^2(x)$

مدينة زويل للعلوم والتكنولوجيا
Zewail City of Science and Technology

Progress Exam # 1 (60 min, 10 pts)

Forward pass

- $h_1 = \tanh(w_1 x_1 + w_3 x_2) = \tanh(v)$
- $h_2 = \tanh(w_2 x_1 + w_4 x_2) = \tanh(u)$
- $o = \tanh(w_5 h_1 + w_6 h_2) = \tanh(z)$
- $E = (o - t)^2$

Updating w_5

$$\frac{\partial E}{\partial w_5} = \frac{\partial E}{\partial o} * \frac{\partial o}{\partial z} * \frac{\partial z}{\partial w_5} = 2(o - t) * \left(1 - tanh^2(z)\right) * h1$$

$$\frac{\partial E}{\partial w_5} = 2h_1\left(\tanh(z) - t\right)\left(1 - tanh^2(z)\right) = 2h_1(o - t)(1 - o^2)$$

$$w_5^+ = w_5 - \eta \frac{\partial E}{\partial w_5}$$

Similarly

$$\frac{\partial E}{\partial w_6} = 2h_2(o - t)(1 - o^2)$$

Updating w1

$$\frac{\partial E}{\partial w_1} = \frac{\partial E}{\partial h_1} * \frac{\partial h_1}{\partial v} * \frac{\partial v}{\partial w_1}$$

$$\frac{\partial E}{\partial h_1} = \frac{\partial E}{\partial o} * \frac{\partial o}{\partial z} * \frac{\partial z}{\partial h_1} = 2w_5(\tanh(z) - t)\,(1 - tanh^2(z))$$

$$\frac{\partial E}{\partial w_1} = 2w_5 x_1\,(\tanh(z) - t)\left(1 - tanh^2(z)\right) * \left(1 - tanh^2(v)\right)$$

$$\frac{\partial E}{\partial w_1} = 2w_5 x_1 \,(\mathrm{o} - t)\,(1 - o^2)(1 - {h_1}^2)$$

Similarly

$$\frac{\partial E}{\partial w_2} = 2w_6 x_1 \,(\mathrm{o} - t)\,(1 - o^2)(1 - {h_2}^2)$$

$$\frac{\partial E}{\partial w_3} = 2w_5 x_2 \,(\mathrm{o} - t)\,(1 - o^2)(1 - {h_1}^2)$$

$$\frac{\partial E}{\partial w_4} = 2w_6 x_2 \,(\mathrm{o} - t)\,(1 - o^2)(1 - {h_2}^2)$$