Name: ______________________

Training Neural Networks with Regularization

8. (8 points) In this problem we will investigate regularization for neural networks.

Kim constructs a fully connected neural network with L=2 layers using mean squared error (MSE) loss and ReLU activation functions for the hidden layer, and a linear activation for the output layer. The network is trained with a gradient descent algorithm on a data set of n points $\{(x^{(1)}, y^{(1)}), ..., (x^{(n)}, y^{(n)})\}$.

Recall that the update rule for weights W^1 can be specified in terms of step size η and the gradient of the loss function with respect to weights W^1. This gradient can be expressed in terms of the activations A^l, weights W^l, pre-activations Z^l, and partials $\frac{\partial L}{\partial A^2}$, $\frac{\partial A^l}{\partial Z^l}$, for l = 1,2:

$$W^1 := W^1 - \eta \sum_{i=1}^{n} \frac{\partial L(h(x^{(i)}; W), y^{(i)})}{\partial W^1},$$

where $h(\cdot)$ is the input-output mapping implemented by the entire neural network, and

$$\frac{\partial L}{\partial W^1} = \frac{\partial Z^1}{\partial W^1} \cdot \frac{\partial A^1}{\partial Z^1} \cdot W^2 \cdot \frac{\partial A^2}{\partial Z^2} \cdot \frac{\partial L}{\partial A^2}.$$

(a) Derive a new update rule for weights W^1 which also penalizes the sum of squared values of all individual weights in the network:

$$L^{new} = L(h(x^{(i)}; W), y^{(i)}) + \lambda ||W||^2$$

where λ denotes the regularization trade-off parameter. You can express the new update rule as follows:

$$W^1 := \alpha W^1 - \eta \sum_{i=1}^{n} \frac{\partial L(h(x^{(i)}; W), y^{(i)})}{\partial W^1}$$

where $L(\cdot)$ represents the previous prediction error loss.

What is the value of α in terms of λ and η?

Name: ______________________________

(b) Explain how this new update rule helps the neural network reduce overfitting to the data.

(c) Given that we are training a neural network with gradient descent, what happens when we increase the regularization trade-off parameter λ too much, while holding the step size η fixed?