Generalized equations for finding the weight and bias gradients

$$\frac{dJ}{dW_2} = \frac{dJ}{dA_2} \frac{dA_2}{dZ_2} \frac{dZ_2}{dW_2}$$

$$\frac{dJ}{db_2} = \frac{dJ}{dA_2} \frac{dA_2}{dZ_2} \frac{dZ_2}{db_2}$$

$$\frac{dJ}{dW_1} = \frac{dJ}{dA_2} \frac{dA_2}{dZ_2} \frac{dZ_2}{dA_1} \frac{dA_1}{dZ_1} \frac{dZ_1}{dW_1}$$

$$\frac{dJ}{db_1} = \frac{dJ}{dA_2} \frac{dA_2}{dZ_2} \frac{dZ_2}{dA_1} \frac{dA_1}{dZ_1} \frac{dZ_1}{db_1}$$

Update Rules

$$W_\ell := W_\ell - \alpha \frac{\partial J}{\partial W_\ell}$$

$$b_\ell := b_\ell - \alpha \frac{\partial J}{\partial b}$$

$$J = (a_2 - y)^2$$

$$\frac{\partial J}{\partial W_2} = a_1^T \cdot 2(a_2 - y) \cdot z_2$$

$$\frac{\partial J}{\partial b_2} = 2(a_2 - y)$$

$$\frac{\partial J}{\partial W_1} = X^T \cdot 2(a_2 - y) \cdot z_2 \cdot W_2^T \cdot g'(a_1)$$

$$\frac{\partial J}{\partial b_1} = 2(a_2 - y) \cdot z_2 \cdot W_2^T \cdot g'(a_1)$$

Now that we have our gradients, we're ready to train.

Step 1: Forward Pass

$$z_1 = W_1 x + b_1$$
$$a_1 = \sigma(z_1)$$

$$z_2 = W_2 a_1 + b_2$$
$$a_2 = \sigma(z_2)$$

Step 2: Back Propagation

Update weights and biases based on the loss, learning rate, and gradients. Do this until the error converges.

$$W_\ell := W_\ell - \alpha \frac{\partial J}{\partial W_\ell}$$

where α is learning rate

$$b_\ell := b_\ell - \alpha \frac{\partial J}{\partial b}$$

Step 3: Test Model

Use the optimized weights and biases on the x_test data and see how the outputs compare to the y_test data.

In my case, I ended with a loss of 79 as mentioned in question 2 of the homework.

How Is This Update Different From Log Loss?

This update rule is very similar. There are minor differences in the gradients.

For log loss update rules:

$$\frac{\partial J}{\partial w_2} = (a_2 - y) \cdot a_1^T$$

$$\frac{\partial J}{\partial b_2} = (a_2 - y)$$

$$\frac{\partial J}{\partial w_1} = (a_2 - y) \cdot w_2 \cdot g'(z_2) \cdot X$$

$$\frac{\partial J}{\partial b_1} = (a_2 - y) \cdot w_2 \cdot g'(z_1)$$

The binary cross entropy cost function is used when doing binary classification problems.

As you can see above, the main differences between the update rules is that the cost function is different. When the cost function changes, all of the partial derivative and gradients will change too.