QUESTION 1

Initialization of weight and bias for two layers:
$w_1, w_2,\ b_1, b_2$

Layer 1 output $z_1 = w_1 . x + b_1$

activation function $a_1 = g(z_1) = \text{sigmoid}(z_1)$

Layer 2 output = predicted value = z_2

$$z_2 = w_2 a_1 + b_2$$

prediction output function $a_2 = g(z_2) = z_2$
(regression use identity function)

Loss function: $L = \frac{1}{2m}(\hat{y} - y)^2$ (MSE for m data points)

Update rule: $w_i := w_i - \alpha \frac{\partial L}{\partial w_i}$ } in layer i th

$b_i := b_i - \alpha \frac{\partial L}{\partial b_i}$ } α: learning rate

$\boxed{\text{Layer 2}}$

$$\frac{\partial L}{\partial w_2} = \frac{1}{2m} \cdot \frac{\partial (z_2 - y)^2}{\partial w_2}$$

$$= \frac{1}{2m} \cdot 2(z_2 - y) \cdot \frac{\partial (z_2 - y)}{\partial w_2}$$

$$= \frac{1}{m}(z_2 - y)\left(\frac{\partial z_2}{\partial w_2} - \frac{\partial y}{\partial w_2}\right)$$

we have $z_2 = w a_1 + b_2$
$\rightarrow \frac{\partial z_2}{\partial w_2} = a_1$

$z_2 = a_2$; $y = \text{const} \Rightarrow dy = 0$

$$\frac{\partial L}{\partial w_2} = \frac{1}{m}(a_2 - y)\, a_1$$

$$\frac{\partial L}{\partial b_2} = \frac{1}{m}(a_2 - y) \cdot \frac{\partial z_2}{\partial b_2} = \frac{1}{m}(a_2 - y)$$

$\boxed{\text{Layer 1}}$ using chain rule :

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial a_2} \cdot \frac{\partial a_2}{\partial z_2} \cdot \frac{\partial z_2}{\partial a_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z_1}{\partial w_1}$$

$$= \frac{1}{m}(a_2 - y) \cdot 1 \cdot w_2 \cdot g'(z_1) \cdot x$$

$g(z_1)$ is a sigmoid function

$$\rightarrow g'(z_1) = g(z_1)(1 - g(z_1)) = a_1(1 - a_1)$$

$$\rightarrow \frac{\partial L}{\partial w_1} = \frac{1}{m}(a_2 - y) \cdot w_2 \cdot a_1(1 - a_1) \cdot x$$

similarly $\frac{\partial L}{\partial b_1} = \frac{1}{m}(a_2 - y) \cdot w_2 \cdot a_1(1 - a_1)$

Model training and compare to binary classification using log loss

Parameters will be initialized, and forward prop will be used to calculated output of every layer. After that, backpropagation will calculate the gradient based on loss function and learning rate to update the parameters until the model converges - parameters are optimal values.

For the update rule, MSE loss depends on number of data points m while log loss does not. Both update rules depends on activation function used in hidden layer.