To train a 2-layer neural network (1 hidden layer and 1 output layer) using backpropagation for regression with the (MSE) loss, first we need to setup the structure for the neural network: Inputs = x, hidden layer = W1 for weights and b1 for biases (sigmoid is the activation function here), output layer W2 for weights and b2 for biases (output just single number, no activation function)

2 - forward Propagation
we calculate the output of the network based on the input.

Hidden Layer: z1 = W1 * x + b1 then the results z1 go in the sigmoid function $a1 = \sigma(z1) = 1/(1 + e^{-z1})$

Output layer: y(pred) = W2 * a1 + b1

3- MSE
Loss (MSE)= ½ (y(pred)−y)^2

4- we update W1, W2, b1, b2 to minimize the loss (Backpropagation)
Using the chain rule we will compute the gradients layer

For the Ouput layer, the error is different between the predicted output and the actual target $\delta 2 = y(pred) - y$,
then now we can compute how this error affects the weights W2 and biases b2:

Gradient w.r.t w2:

$$\partial Loss/\partial W2 = \delta 2 \cdot a1$$

Gradient w.r.t b2

$$\partial Loss/\partial b2 = \delta 2$$

Now we gotta propagate this error back to the hidden layer $\delta 1 = \delta 2 \cdot W2 \cdot \sigma'(z1)$
Now we get sigma derivate of z1 $\sigma'(z1) = a1(1 - a1)$

then we compute how this affects the weights w1 and b1
Gradient w.r.t w2:

$$\partial Loss/\partial W1 = \delta 1 \cdot x$$

Gradient w.r.t b2

$$\partial Loss/\partial b1 = \delta 1$$

Now we have the gradients, now we update W and b using gradient descent
$W2 \leftarrow W2 - \eta \frac{\partial Loss}{\partial W2}$, $b1 \leftarrow b2 - \eta \frac{\partial Loss}{\partial b2}$

$W1 \leftarrow W1 - \eta \frac{\partial Loss}{\partial W1}$, $b1 \leftarrow b2 - \eta \frac{\partial Loss}{\partial b1}$
Where η (eta) is the learning rate that controls the step size of the updates.

Explain briefly how this is different from the update rule for the network trained for binary classification using log loss.

In binary classification with log loss instead of using MSE we use log loss to measure the difference between the predicted probability and the actual binary label (0 or 1).
For regression with MSE: The output is a continuous number, and we use MSE to calculate the error.
For binary classification with log loss: The output is a probability (0 or 1), and we use log loss (cross-entropy) to measure the error.