Let L be the loss function.

Z_l is the pre-activation at layer l

B_l is the parameter matrix (weights and biases) for layer l.

Then the gradient of loss with respect to parameters at layer l

$$\frac{dL}{dB_l} = \frac{dL}{dZ_l} \cdot \frac{dZ_l}{dB_l}$$

Chain rule for gradients state that

$$\frac{dL}{dB_l} = \frac{dL}{dZ_L} \cdot \frac{dZ_L}{dZ_{L-1}} \cdot \frac{dZ_{L-1}}{dZ_{L-2}} \cdots \frac{dZ_{l+1}}{dZ_l} \cdot \frac{dZ_l}{dB_l}$$

Considering all weights are initialized to 0

$\frac{dZ_l}{dB_l} = 0$ when multiplied in the chain rule will result to 0 for all layers following the gradient descent.

For the last layer L, the error is directly dependent on the final loss.

Therefore, $\frac{dL}{dB_L} \neq 0$ for the final layer.

The chain rule for gradient descent will not equate to zero as $\frac{dL}{dB_L} \neq 0$.

Therefore, during back propagation from final loss the parameters will update.