Input:

Training data: $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$

Current weights: $W_k^2, W_k^3, \ldots, W_k^L$

Current biases: $b_k^2, b_k^3, \ldots, b_k^L$

Learning rate: $\eta \in \mathbb{R}$

Output:

Updated weights: $W_{k+1}^2, W_{k+1}^3, \ldots, W_{k+1}^L$

Updated biases: $b_{k+1}^2, b_{k+1}^3, \ldots, b_{k+1}^L$

for $i \leftarrow 1$ **to** n **do**

Forward pass:

begin

$a^1 = X_i$

for $\ell \leftarrow 2$ **to** L **do**

$z^\ell = W_k^\ell a^{\ell-1} + b^\ell$

$a^\ell = \tilde{\sigma}(z^\ell)$

end

end

Compute errors (backpropagation):

begin

$\delta^L = [a^L(X_i) - Y_i] \odot \tilde{\sigma}'(z^L)$

$\frac{\partial C_i}{\partial W_k^L} = \delta^L (a^{L-1})^T$

$\frac{\partial C_i}{\partial b_k^L} = \delta^L$

for $\ell \leftarrow L-1, L-2, \ldots$ **to** 2 **do**

$\delta^\ell = \left[\left(W^{\ell+1} \right)^T \delta^{\ell+1} \right] \odot \tilde{\sigma}'(z^\ell)$

$\frac{\partial C_i}{\partial W_k^\ell} = \delta^\ell \left(a^{\ell-1} \right)^T$

$\frac{\partial C_i}{\partial b_k^\ell} = \delta^\ell$

end

end

end

Compute gradients:

begin

for $\ell \leftarrow 2$ **to** L **do**

$\frac{\partial C}{\partial W_k^\ell} = \frac{1}{n} \sum_{i=1}^n \frac{\partial C_i}{\partial W_k^\ell}$

$\frac{\partial C}{\partial b_k^\ell} = \frac{1}{n} \sum_{i=1}^n \frac{\partial C_i}{\partial b_k^\ell}$

end

end

Update weights and biases (gradient descent):

begin

for $\ell \leftarrow 2$ **to** L **do**

$W_{k+1}^\ell = W_k^\ell - \eta \frac{\partial C}{\partial W_k^\ell}$

$b_{k+1}^\ell = b_k^\ell - \eta \frac{\partial C}{\partial b_k^\ell}$

end

end