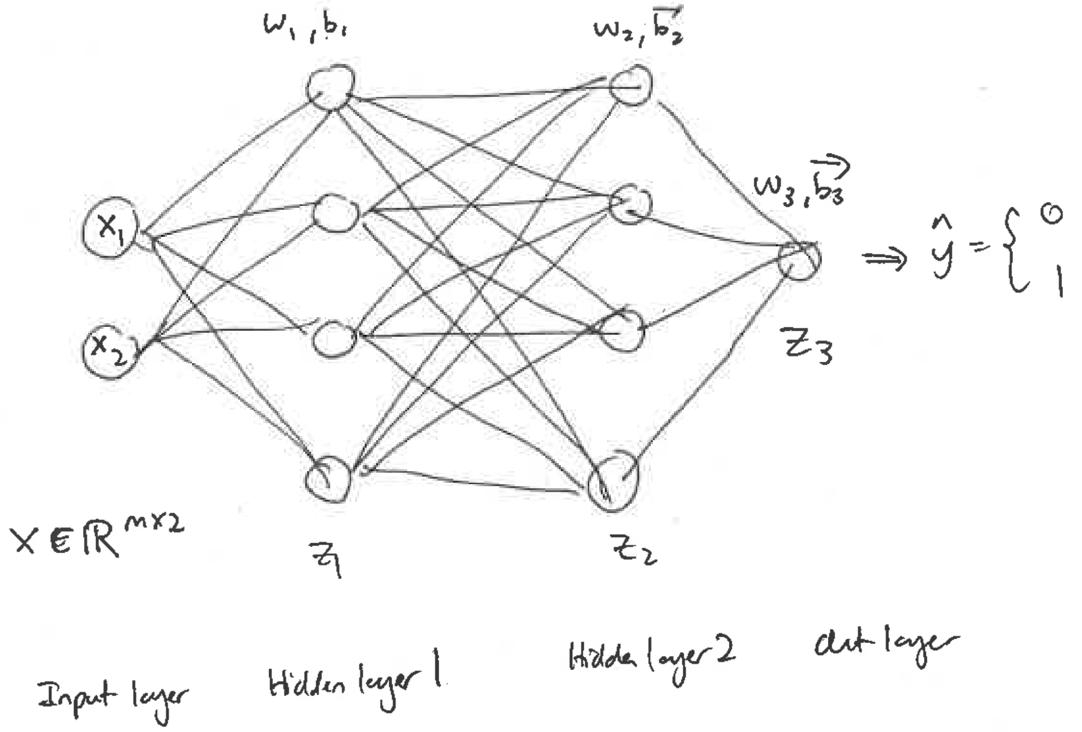

<u>Feedforward</u>:

$$Z_1 = X \cdot W_1 + \vec{b_1}$$
$$A_1 = g_1(Z_1)$$; where g_1 = activation function

$$Z_2 = A_1 \cdot W_2 + \vec{b_2}$$
$$A_2 = g_2(Z_2)$$; where g_2 = activation function

$$Z_3 = A_2 \cdot W_3 + \vec{b_3}$$
$$A_3 = g_3(Z_3) = \hat{y}$$; where g_3 = activation function

$g_1, g_2 \rightarrow$ ReLU activation

$g_3 \rightarrow$ Sigmoid activation

<u>Loss</u>: Cross-entropy, $\mathcal{L} = \frac{-1}{m} \sum_{i=0}^{m} \left(y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}) \right)$

$$\frac{\partial \mathcal{L}}{\partial \hat{y}} = \frac{-y}{\hat{y}} + \frac{1-y}{1-\hat{y}}$$

Activation functions:

$$\mathrm{ReLU}(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{d}{dx}\mathrm{ReLU}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

$$\begin{aligned}
\frac{d}{dx}\sigma(x) = \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right) &= -(1+e^{-x})^{-2}\cdot(-e^{-x}) \\
&= \frac{e^{-x}}{(1+e^{-x})^2} \\
&= \frac{1}{1+e^{-x}}\cdot\frac{e^{-x}}{1+e^{-x}} \\
&= \frac{1}{(1+e^{-x})}\cdot\frac{(1+e^{-x})-1}{(1+e^{-x})} \\
&= \frac{1}{1+e^{-x}}\cdot\left(1-\frac{1}{1+e^{-x}}\right) \\
&= \sigma(x)\cdot(1-\sigma(x))
\end{aligned}$$

Backpropagation:

$$\begin{aligned}
\frac{\partial L}{\partial z_3} = \frac{\partial L}{\partial \hat{y}}\cdot\frac{\partial \hat{y}}{\partial z_3} &= \frac{\partial L}{\partial \hat{y}}\cdot\frac{\partial \sigma(z_3)}{\partial z_3} \\
&= \left(\frac{-y}{\hat{y}} + \frac{1-y}{1-\hat{y}}\right)\cdot\left(\hat{y}(1-\hat{y})\right) \\
&= \frac{-y(1-\hat{y}) + (1-y)\hat{y}}{\hat{y}(1-\hat{y})}\cdot\hat{y}(1-\hat{y}) \\
&= \hat{y} + y\cdot y + \hat{y} - y\hat{y} \\
&= \hat{y} - y
\end{aligned}$$

$$\frac{\partial L}{\partial W_3} = \frac{\partial L}{\partial Z_3} \cdot \frac{\partial Z_3}{\partial W_3} = \frac{\partial L}{\partial Z_3} \cdot \frac{\partial (A_3 \cdot W_3 + b_3)}{\partial W_3}$$

$$\therefore \frac{\partial L}{\partial W_3} = \hat{y} - y \cdot A_3$$

$$\frac{\partial L}{\partial b_3} = \frac{\partial L}{\partial Z_3} \cdot \frac{\partial Z_3}{\partial b_3} = \hat{y} - y \quad ; \frac{\partial Z_3}{\partial b_3} = 1$$

$$\frac{\partial L}{\partial Z_2} = \frac{\partial L}{\partial A_2} \cdot \frac{\partial A_2}{\partial Z_2} = \frac{\partial L}{\partial Z_3} \cdot \frac{\partial Z_3}{\partial A_2} \cdot \frac{\partial A_2}{\partial Z_2}$$

$$= \hat{y} - y \cdot W_3 \cdot g_3'(Z_2)$$

$$\frac{\partial L}{\partial W_2} = \frac{\partial L}{\partial Z_2} \cdot \frac{\partial Z_2}{\partial W_2} = \frac{\partial L}{\partial Z_2} \cdot A_1$$

$$\frac{\partial L}{\partial b_2} = \frac{\partial L}{\partial Z_2} \cdot \frac{\partial Z_2}{\partial b_2} = \frac{\partial L}{\partial Z_2} \cdot 1$$

$$\frac{\partial L}{\partial Z_1} = \frac{\partial L}{\partial A_1} \cdot \frac{\partial A_1}{\partial Z_1} = \frac{\partial L}{\partial Z_2} \cdot \frac{\partial Z_2}{\partial A_1} \cdot \frac{\partial A_1}{\partial Z_1}$$

$$= \frac{\partial L}{\partial Z_2} \cdot W_2 \cdot g_1'(Z_1)$$

$$\frac{\partial L}{\partial W_1} = \frac{\partial L}{\partial Z_1} \cdot \frac{\partial Z_1}{\partial W_1} = \frac{\partial L}{\partial Z_1} \cdot X$$

$$\frac{\partial L}{\partial b_1} = \frac{\partial L}{\partial Z_1} \cdot \frac{\partial Z_1}{\partial b_1} = \frac{\partial L}{\partial Z_1} \cdot 1$$

Update parameters: $\theta := \theta - \alpha \frac{\partial L}{\partial \theta}$; where α = learning rate

$$W_3 = W_3 - \alpha \frac{\partial L}{\partial W_3} \qquad W_2 = W_2 - \alpha \frac{\partial L}{\partial W_2} \qquad W_1 = W_1 - \alpha \frac{\partial L}{\partial W_1}$$

$$b_3 = b_3 - \alpha \frac{\partial L}{\partial b_3} \qquad b_2 = b_2 - \alpha \frac{\partial L}{\partial b_2} \qquad b_1 = b_1 - \alpha \frac{\partial L}{\partial b_1}$$