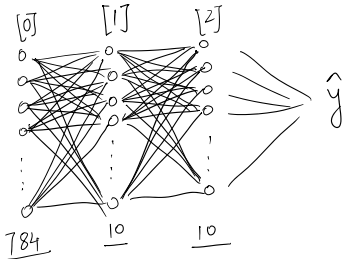


784 m training images
28x28

$$X = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(m)} \end{bmatrix}^T = \begin{bmatrix} | & | & \dots & | \\ x^{(1)} & x^{(2)} & \dots & x^{(m)} \\ | & | & \dots & | \end{bmatrix}$$

784 \Rightarrow 0,1,2,...,9
28x28 10 classes

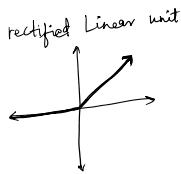
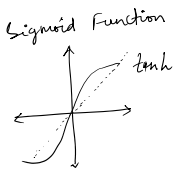


① Forward Propagation

$$A^{[0]} = X \quad (784 \times m)$$

$$Z^{[1]} = W^{[0]} A^{[0]} + b^{[1]} \quad \begin{matrix} 10 \times m & 10 \times 784 & 784 \times m & 10 \times 1 \Rightarrow 10 \times m \end{matrix}$$

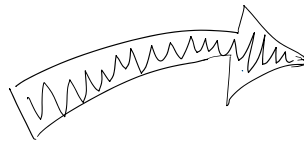
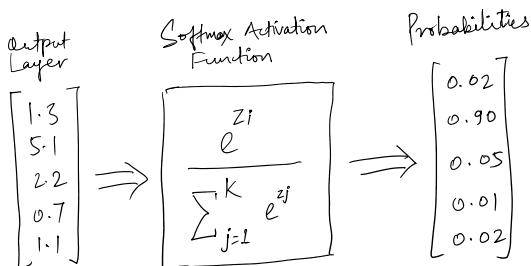
$$A^{[1]} = g(Z^{[1]}) = \text{ReLU}(Z^{[1]})$$



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

$$Z^{[2]} = W^{[1]} A^{[1]} + b^{[2]} \quad \begin{matrix} 10 \times m & 10 \times 10 & 10 \times m & 10 \times 1 \Rightarrow 10 \times m \end{matrix}$$

$$A^{[2]} = \text{Softmax}(Z^{[2]})$$



② Back Propagation

$$dZ^{[2]} = A^{[2]} - Y \quad \begin{matrix} 10 \times m & 10 \times m & 10 \times m \end{matrix}$$

$$dW^{[2]} = \frac{1}{m} dZ^{[2]} A^{[1]T} \quad \begin{matrix} 10 \times 10 & 10 \times m & m \times 10 \end{matrix}$$

$$db^{[2]} = \frac{1}{m} \sum dZ^{[2]} \quad \begin{matrix} 10 \times 1 & 10 \times 1 \end{matrix}$$

$$dZ^{[1]} = W^{[2]T} dZ^{[2]} \cdot g'(Z^{[1]}) \quad \begin{matrix} 10 \times m & 10 \times 10 & 10 \times m & 10 \times m \end{matrix}$$

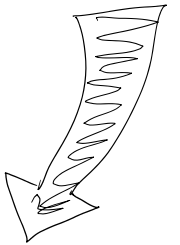
$$dW^{[1]} = \frac{1}{m} dZ^{[1]} X^T \quad \begin{matrix} 10 \times 784 & 10 \times m & m \times 784 \end{matrix}$$

$$db^{[1]} = \frac{1}{m} \sum dZ^{[1]} \quad \begin{matrix} 10 \times 1 & 10 \times 1 \end{matrix}$$

0.01
0.02
0.05
0.03
0.9
...

One-hot encode
ie. $y = 4$

\Rightarrow $\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$



③ Update Parameters

$$W^{[1]} := W^{[1]} - \alpha dW^{[1]}$$

$$b^{[1]} := b^{[1]} - \alpha db^{[1]}$$

$$W^{[2]} := W^{[2]} - \alpha dW^{[2]}$$

$$b^{[2]} := b^{[2]} - \alpha db^{[2]}$$

α learning rate