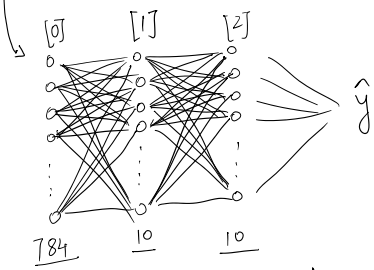
784 (28 × 28) — m training images

$$X = \begin{bmatrix} — x^{(1)} — \\ — x^{(2)} — \\ \vdots \\ — x^{(n)} — \end{bmatrix}^T = \begin{bmatrix} | & | & & | \\ x^{(1)} & x^{(2)} & \cdots & x^{(n)} \\ | & | & & | \end{bmatrix}$$

784 (28 × 28) ⇒ 0, 1, 2 .... 9 — 10 classes



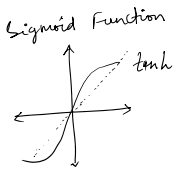
① Forward Propagation

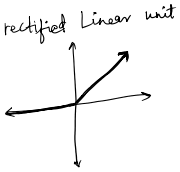
$$A^{[0]} = X \qquad (784 \times m)$$

$$Z^{[1]} = W^{[1]} A^{[0]} + b^{[1]}$$

$(10 \times m) \quad (10 \times 784) \quad (784 \times m) \quad (10 \times 1) \Rightarrow (10 \times m)$

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



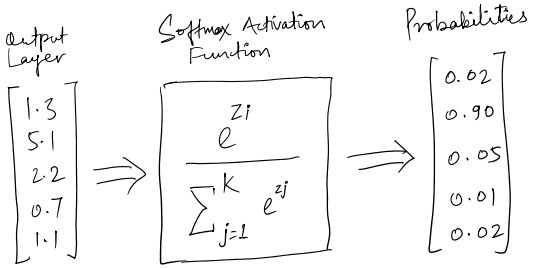


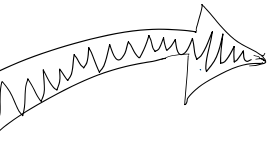
$$\text{ReLU}(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{if } x \le 0 \end{cases}$$

$$Z^{[2]} = W^{[2]} A^{[1]} + b^{[2]}$$

$(10 \times m) \quad (10 \times 10) \quad (10 \times m) \quad (10 \times 1) \Rightarrow (10 \times m)$

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





② Back Propagation

One-hot encode
ie: $y = 4$
$\Rightarrow$

$$dZ^{[2]} = A^{[2]} - Y \qquad (10 \times m) \quad (10 \times m) \quad (10 \times m)$$

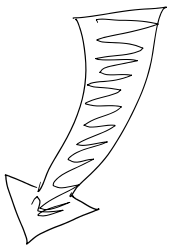
$$dW^{[2]} = \frac{1}{m} dZ^{[2]} A^{[1]T} \qquad (10 \times 10) \quad (10 \times m) \quad (m \times 10)$$

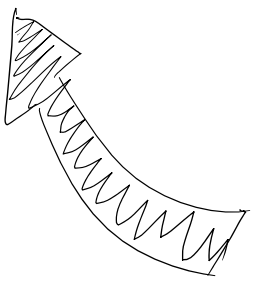
$$db^{[2]} = \frac{1}{m} \sum dZ^{[2]} \qquad (10 \times 1) \quad (10 \times 1)$$

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

$$dW^{[1]} = \frac{1}{m} dZ^{[1]} X^T \qquad (10 \times 784) \quad (10 \times m) \quad (m \times 784)$$

$$db = \frac{1}{m} \sum dZ^{[1]} \qquad (10 \times 1) \quad (10 \times 1)$$





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