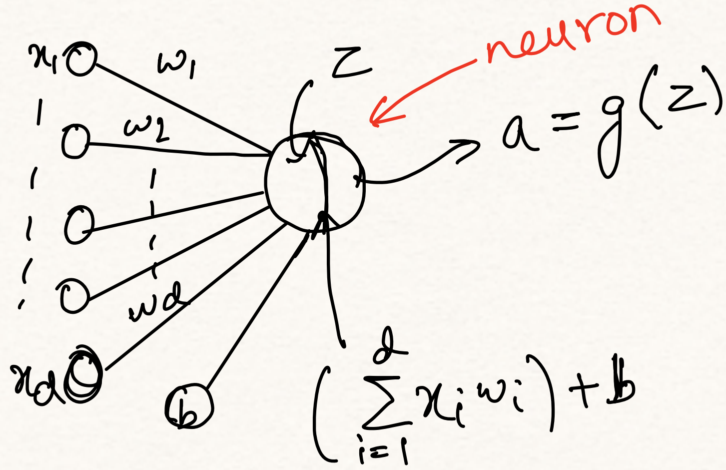
$x \in \mathbb{R}^d$



For logistic reg.

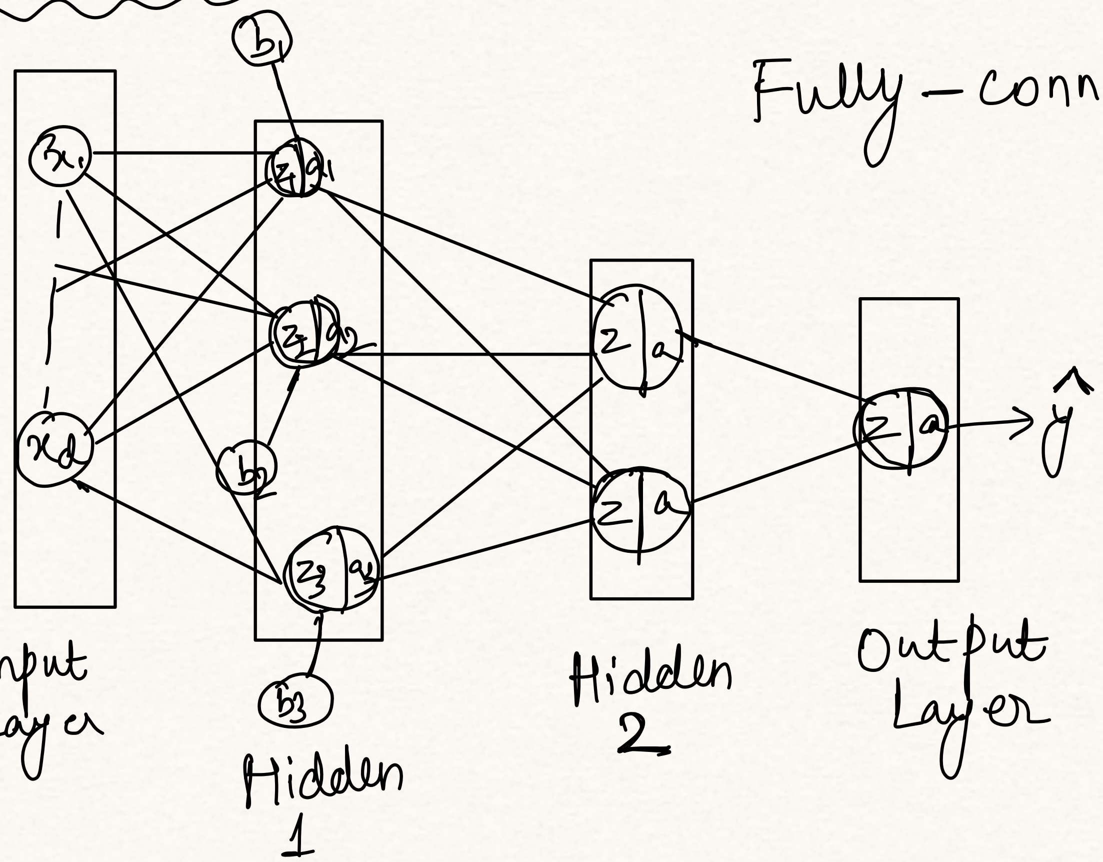
$$g(z) = \frac{1}{1+e^{-z}}$$

activation function

$$z = x^T w + b$$

Neural Network

Fully-connected



y

Then calculate loss for $\hat{y}$.

Terminologies

layer #

$w_{ij}^{[l]}$ — connection

$w^{[l]}$ is weight matrix

$b_i^{[l]}$ — bias

$z_i^{[l]}$ — $w\cdot(\ \ )+b$

$a_i^{[l]}$ — $g(z)$

x — $a^{[0]}$

$$\text{\# of neurons in the } l^{th} \text{ layer}\left\{\underbrace{\begin{bmatrix} & & \\ & W^{[l]} & \\ & & \end{bmatrix}}_{\text{\# of neurons in the } (l-1)^{th} \text{ layer.}}\right.$$

$$z_1^{[1]} = \sum_j w_{1j}^{[1]} a_j^{[0]} + b_1^{[1]}$$

$$z_2^{[1]} = w_2^{[1]T} a^{[0]} + b_2^{[1]}$$

$$a_1^{[1]} = g\left(z_1^{[1]}\right)$$

$$\boxed{\underbrace{Z^{[1]}}_{3} = \underbrace{W^{[1]}}_{3\times d}\,\underbrace{a^{[0]}}_{d} + \underbrace{b^{[1]}}_{3}}$$

$$a^{[1]} = g\left(z^{[1]}\right)$$

$$z^{[2]} = W^{[2]} a^{[1]} + b^{[2]}$$

$$a^{[2]} = g\left(z^{[2]}\right)$$

$$z^{[3]} = W^{[3]} a^{[2]} + b^{[3]}$$

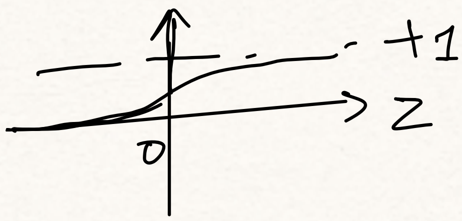
$$a^{[3]} = g\left(z^{[3]}\right)$$

$$\vdots$$

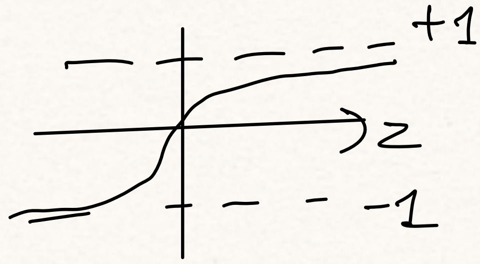
What if $g(z) = z$, $a = z$

$$
\begin{aligned}
a^{[3]} &= W^{[3]} a^{[2]} \\
&= W^{[3]} z^{[2]} \\
&= W^{[3]} W^{[2]} a^{[1]} \\
&= W^{[3]} W^{[2]} z^{[1]} = W^{[3]} \cdot W^{[2]} \cdot W^{[1]} \cdot x \\
&= \tilde{W} \cdot x
\end{aligned}
$$

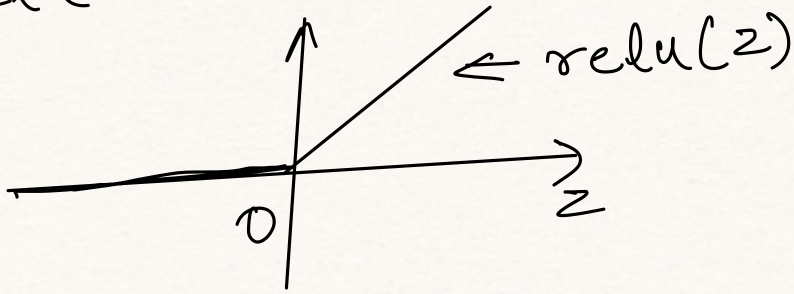
$$g(z) = \frac{1}{1+e^{-z}}$$



$$g(z) = \tanh(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$$



$$g(z) = \text{relu} = \max(z, 0)$$



$$a^{[0]} = x^{(i)}$$
$$z^{[1]} = W^{[1]} a^{[0]} + b^{[1]}$$
$$a^{[1]} = g(z^{[1]})$$
$$z^{[2]} = W^{[2]} a^{[1]} + b^{[2]}$$
$$a^{[2]} = g(z^{[2]})$$
$$\vdots$$
$$a^{[L]} = g(z^{[L]})$$
$$\hat{y}^{(i)} = a^{[L]}$$
$$\mathcal{L}(y^{(i)}, \hat{y}^{(i)}) = -\left[y^{(i)} \log \hat{y}^{(i)} + (1-y^{(i)}) \log(1-\hat{y}^{(i)})\right]$$

For l in $1, 2, --, L$

$$W^{[l]} := W^{[l]} - \alpha \cdot \frac{\partial \mathcal{L}}{\partial W^{[l]}}$$

$$b^{[l]} := b^{[l]} - \alpha \cdot \frac{\partial \mathcal{L}}{\partial b^{[l]}}$$

not to be initialized 0.

Initialize w, b

For t in $1, ---$

$$\theta := \theta - \alpha (\quad)$$

Xavier/He init.

$$w_{ij}^{[l]} \sim N\left(0, \sqrt{\frac{2}{n^{[l]} + n^{[l-1]}}}\right)$$

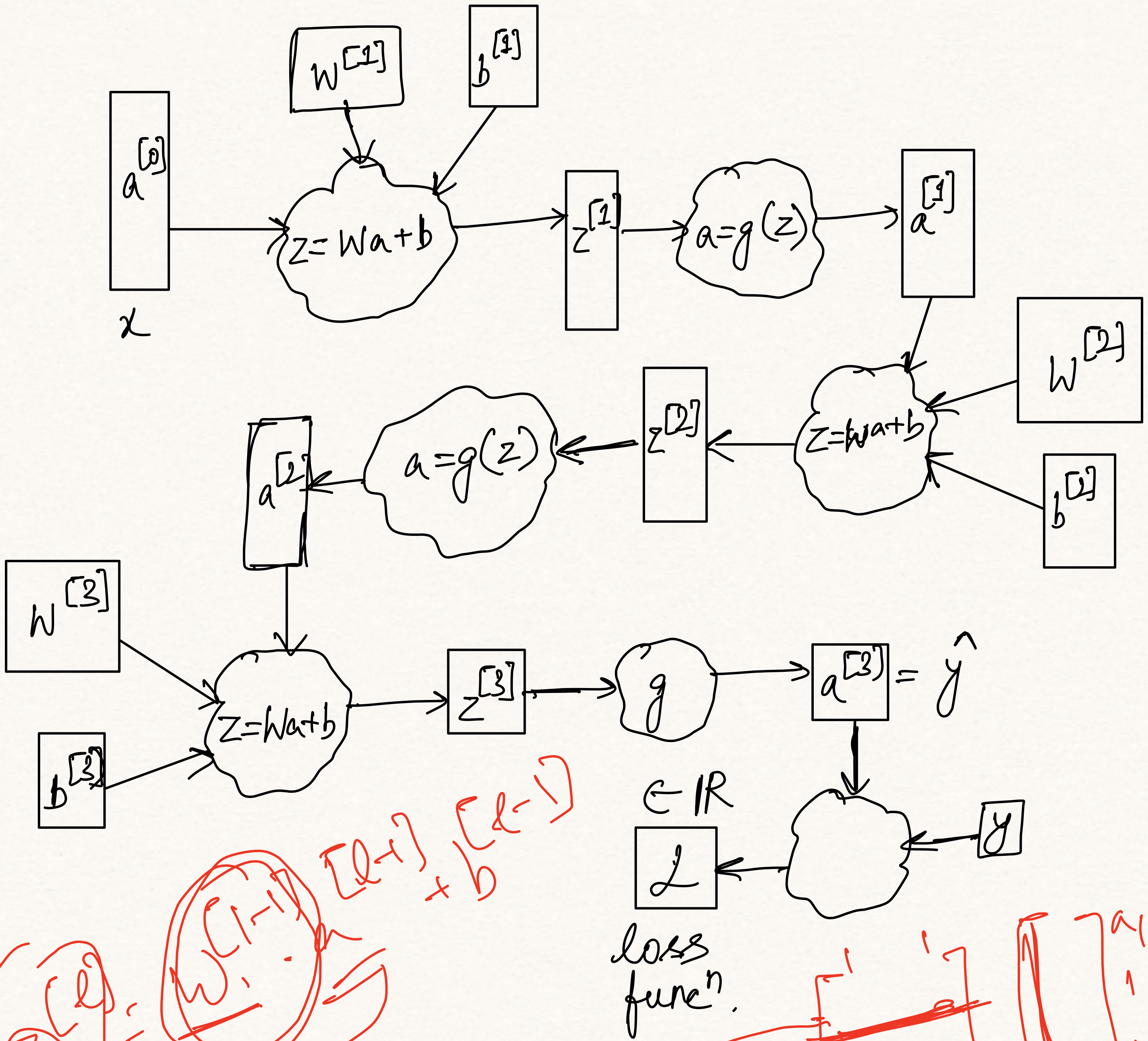
$$W_{ij}^{[l]} \sim U([-0.1, 0.1])$$

For t in $1, 2, \ldots$

For l in $1, 2, \ldots, L$

$$W^{[l]} := W^{[l]} - \alpha \frac{\partial L}{\partial W^{[l]}}$$

$$b^{[l]} := \ldots$$



$$z^{[l]} = W^{[l-1]} \cdot a^{[l-1]} + b^{[l-1]}$$

We are considering binary classification.

$$\frac{\partial \mathcal{L}}{\partial W^{[2]}} = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial W_{11}^{[2]}} & \cdots & \frac{\partial \mathcal{L}}{\partial W_{13}^{[2]}} \\ \vdots & \ddots & \\ \cdots & \cdots & \frac{\partial \mathcal{L}}{\partial W_{33}^{[2]}} \end{bmatrix}$$

$$\frac{\partial \mathcal{L}}{\partial z^{[3]}} = \frac{\partial}{\partial z^{[3]}} \left[-y \log \hat{y} - (1-y) \log(1-\hat{y}) \right]$$

$$= \frac{\partial}{\partial z^{[3]}} \left[-y \log(\sigma(z^{[3]})) - (1-y) \times \log(1-\sigma(z^{[3]})) \right]$$

$$= - -$$

$$= a^{[3]} - y$$

$$\frac{\partial \mathcal{L}}{\partial W_{ij}^{[2]}} = \frac{\partial \mathcal{L}}{\partial a^{[3]}} \cdot \frac{\partial a^{[3]}}{\partial W_{ij}^{[2]}}$$

$$= \frac{\partial \mathcal{L}}{\partial a^{[3]}} \cdot \frac{\partial a^{[3]}}{\partial z^{[3]}} \cdot \frac{\partial z^{[3]}}{\partial W_{ij}^{[2]}}$$

diag[g']

$$= \underbrace{\frac{\partial \mathcal{L}}{\partial a^{[3]}} \cdot \frac{\partial a^{[3]}}{\partial z^{[3]}}}_{(a^{[3]}-y)} \cdot \underbrace{\frac{\partial z^{[3]}}{\partial a^{[2]}}}_{W^{[3]}} \cdot \frac{\partial a^{[2]}}{\partial z^{[2]}} \cdot \frac{\partial z^{[2]}}{\partial W_{ij}^{[2]}}$$

BACKPROPAGATION

$$\begin{array}{rc} & \begin{bmatrix} 0 \\ 0 \\ \vdots \\ a_j^{[1]} \\ \vdots \end{bmatrix} \end{array}$$

i^{th} pos. $\rightarrow$

$$= \underbrace{(a^{[3]}-y)\, W^{[3]} \circ g'(z^{[2]})}_{1\times 2} \underbrace{\begin{bmatrix} 0 \\ \vdots \\ a_j^{[1]} \\ \vdots \\ 0 \end{bmatrix}}_{2\times 1} \leftarrow i$$

$$= \left[(a^{[3]}-y)\, w^{[3]} \circ g'(z^{[2]})\right]_i a_j^{[1]}$$

$$\frac{\partial \mathcal{L}}{\partial W^{[2]}} = (a^{[3]}-y) W^{[3]} \circ g'(z^{[2]})\, {a_j^{[1]}}^T$$

$$f : \mathbb{R}^n \to \mathbb{R}^m$$

$$\frac{\partial f}{\partial x} \in \mathbb{R}^{n \times m}$$

diag. matrix

$W^{[L-1]}$ diag(g')

$$\frac{\partial \mathcal{L}}{\partial W_{ij}^{[2]}} = \underbrace{\frac{\partial \mathcal{L}}{\partial a^{[L]}} \cdot \frac{\partial a^{[L]}}{\partial z^{[L]}} \cdot \frac{\partial z^{[L]}}{\partial a^{[L-1]}}}_{(y - a^{[L]}) \cdot W^{[L]}} \cdot \frac{\partial a^{[L-1]}}{\partial z^{[L-1]}} \cdots \frac{\partial a^{[2]}}{\partial z^{[2]}} \cdot \frac{\partial z^{[2]}}{\partial W_{ij}^{[2]}}$$

$$a_i = g(z_i)$$

$$\frac{\partial a_j}{\partial z_i} = 0 \; ; \; i \neq j$$

$$\frac{\partial a^{[L-1]}}{\partial z^{[L-1]}} = \begin{array}{cc} & z \\ a & \begin{bmatrix} g'() & 0 & 0 & 0 \\ 0 & g'() & & 0 \\ 0 & & \ddots & 0 \\ \vdots & & & g'() \end{bmatrix} \end{array}$$

diagonal matrix

$$z^{[L-1]} = W^{[L-1]} \cdot a^{[L-2]} + b^{[L-1]}$$

$$\boxed{\frac{\partial z^{[L-1]}}{\partial a^{[L-2]}} = W^{[L-1]}}$$

$$\frac{\partial z^{[2]}}{\partial W_{ij}^{[2]}} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ a_j \\ 0 \\ \vdots \end{bmatrix} \leftarrow i\text{th}$$

$$J(w,b) = \sum_{i=1}^{B} \mathcal{L}(y^{(i)}, \hat{y}^{(i)})$$

B: size of minibatch

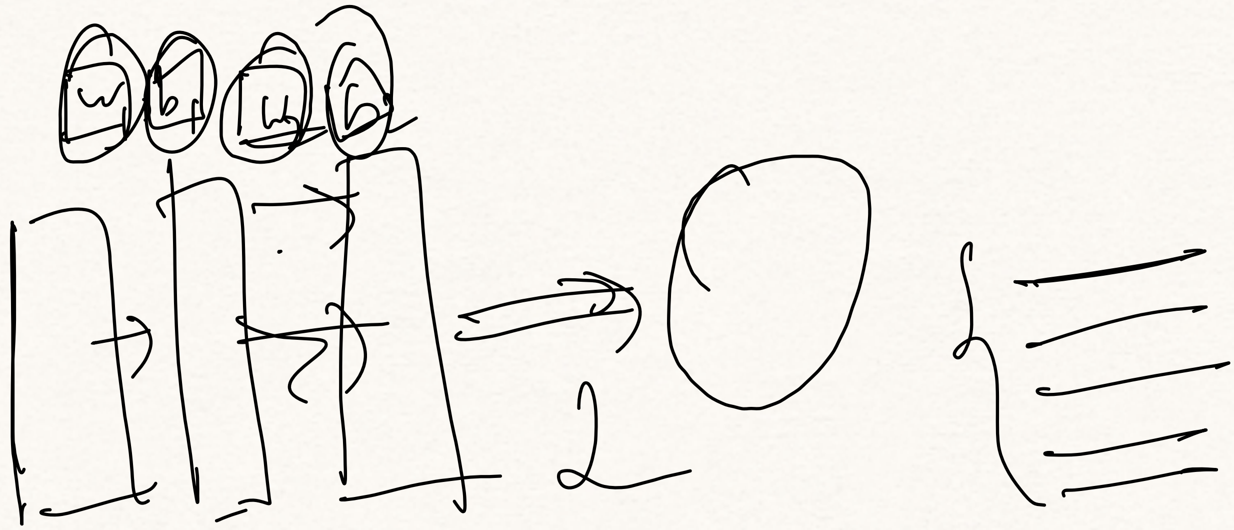
$$\underbrace{z^{[1]}}_{m \times 1} = \underbrace{\underbrace{W^{[1]}}_{m \times d} \underbrace{x^{(i)}}_{d \times 1}}_{m \times 1} + \underbrace{b^{[1]}}_{m \times 1}$$

$$= \underbrace{W^{[1]}}_{m \times d} \cdot \underbrace{[x^{(1)} x^{(2)} \cdots x^{(B)}]}_{d \times B} + \underbrace{b^{[1]}}_{m \times 1}$$

$$= \quad m \times B \quad + \quad m \times 1 \quad (1 \times B)$$

("Broadcasting") $m\{[\,|\,] \rightarrow \underbrace{[\,|\,|\,|\,]}_{B}$ copies of col.

N.N = Learnable feature map
+
Linear model



$$\frac{\partial J}{\partial w_1}, \frac{\partial J}{\partial b_1} \;---$$

$\mathbb{R}^{10}$

$\emptyset$

$$y = \begin{bmatrix} 1 & 0 & --- & - & 0 \end{bmatrix}^T$$

$J = K$ (i)

0 1

$$-\sum_{i=1}^{n} y_i \log y$$