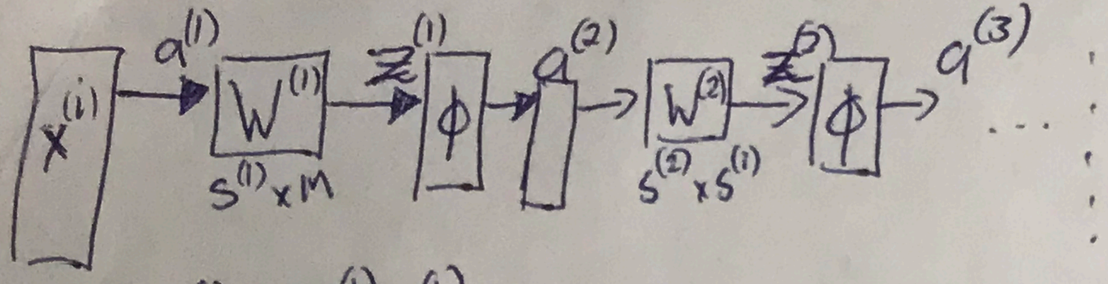
① Recall how a Neural Networks Works



$$X = \begin{bmatrix} — x^{(1)} — \\ — x^{(2)} — \\ \vdots \\ — x^{(M)} — \end{bmatrix}$$

$$A^{(1)} = X^T$$

$$z^{(1)} = W^{(1)} a^{(1)}$$

$$z^{(1)} = \begin{bmatrix} — {}_1W^{(1)} — \\ — {}_2W^{(1)} — \\ \vdots \\ — {}_{s^{(1)}}W^{(1)} — \end{bmatrix} \begin{bmatrix} | \\ a^{(1)} \\ | \end{bmatrix}^{(i)}$$

$$= \begin{bmatrix} {}_1W^{(1)} a^{(1)} \\ {}_2W^{(1)} a^{(1)} \\ \vdots \\ {}_{s^{(1)}}W^{(1)} a^{(1)} \end{bmatrix}$$

$$z^{(1)} = \begin{bmatrix} — {}_1W — \\ \vdots \\ {}_{s^{(1)}}W \end{bmatrix} \begin{bmatrix} | & | & & | \\ [a^{(1)}]^{(1)} & [a^{(1)}]^{(2)} & \cdots & [a^{(1)}]^{(M)} \\ | & | & & | \end{bmatrix}$$

$$z^{(1)} = W^{(1)} \cdot A^{(1)}$$

$$= \begin{bmatrix} | & | & & | \\ [z^{(1)}]^{(1)} & [z^{(1)}]^{(2)} & \cdots & [z^{(1)}]^{(M)} \\ | & | & & | \end{bmatrix}$$

$$A^{(2)} = \phi(z^{(1)})$$

$$= \begin{matrix} | & | & & | \\ \phi(z^{(1)}) & \phi(z^{(1)}) & \cdots & = A^{(2)} \\ | & | & & \end{matrix}$$

$$\begin{matrix} | & | \\ (1) & (2) \end{matrix}$$

$$z^{(2)} = W^{(2)} \cdot A^{(2)}$$

$$\cdots$$

$$z^{(3)} = W^{(3)} \cdot A^{(3)}$$

ETC.

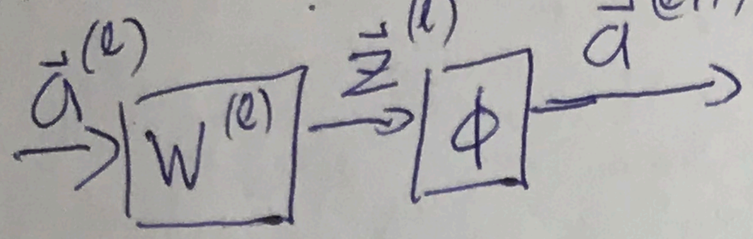
- TO UPDATE $W^{(k)}$, WE NEED GRADIENT OF THE OBJECTIVE FUNCTION, $J(W)$

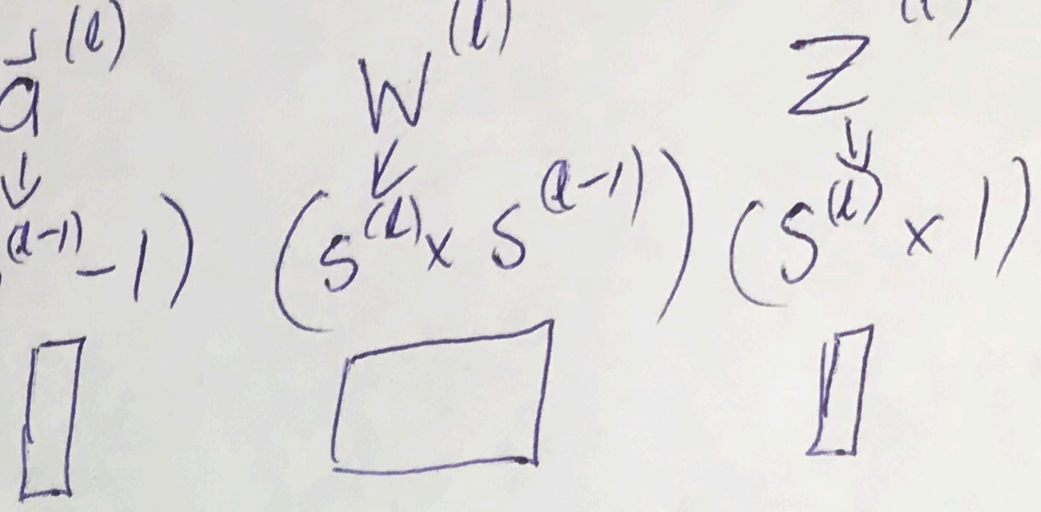
$$W_{ij}^{(k)} \Leftarrow W_{ij}^{(k)} + \frac{\eta \partial J(W)}{\partial W_{ij}^{(k)}}$$

1. REMEMBER WHY THIS DIFFICULT
2. DERIVE $\frac{\partial J}{\partial W}$ FOR ONE INSTANCE $x^{(i)}$
3. DEFINE BACK PROPAGATION
4. DERIVE FAST LINEAR ALGEBRA NOTATION FOR ALL INSTANCES

$$X = \begin{bmatrix} — x^{(1)} — \\ \vdots \\ x^{(M)} \end{bmatrix}$$

- CAN WE WRITE DOWN THE FUNCTIONS OF THE NETWORK AS A SERIES OF COMBINED EQUATIONS?





$$z^{(\ell)} = W^{(\ell)} a^{(\ell)}$$

$$a^{(\ell+1)} = \phi(z^{(\ell)}) = \phi(W^{(\ell)} a^{(\ell)})$$

$$= \phi(W^{(\ell)} \phi(W^{(\ell-1)} a^{(\ell-1)}))$$

$$= \phi(W^{(\ell)} \phi(W^{(\ell-1)} \phi(W^{(\ell-2)} a^{(\ell-2)})))$$

IF WE WANT TO KNOW DERIVATIVES, WE NEED TO USE CHAIN RULE!!

- WHY DO WE NEED DERIVATIVE?

WE NEED GRADIENT OF OBJECTIVE FUNCTION WITH RESPECT TO EACH $W^{(\ell)}$

$$J(W) = \sum_{k=1}^{M} | y^{(k)} - [a^{(L)}]^{(k)} |^2$$

TO GET OUR UPDATE EQUATIONS, WE NEED

$$W^{(\ell)} \leftarrow W^{(\ell)} + \eta \frac{\partial J(W)}{\partial W^{(\ell)}}$$

OR

$$w_{ij}^{(\ell)} \leftarrow w_{ij}^{(\ell)} + \eta \frac{\partial J(W)}{\partial w_{ij}^{(\ell)}}$$

GETTING THIS IS AN EXERCISE IN USING CHAIN RULE..!!

SO LET'S REVIEW THE CHAIN RULE

$$\frac{\partial}{\partial x} f(g(x)) = \frac{\partial f}{\partial g}\frac{\partial g}{\partial x}$$

$$f(x) = \ln(\underbrace{1+x^2}_{g}) \qquad \frac{\partial}{\partial g}\ln(g) = \frac{1}{g}$$

$$\frac{\partial f}{\partial x} = \frac{1}{1+x^2}\frac{\partial}{\partial x}(1+x^2)$$

$$= \frac{1}{1+x^2}\,2x = \frac{2x}{1+x^2}$$

AND IT WORKS CHAINED AS MANY TIMES AS WE WANT!

$$f = (1+\ln(x^3))^2$$

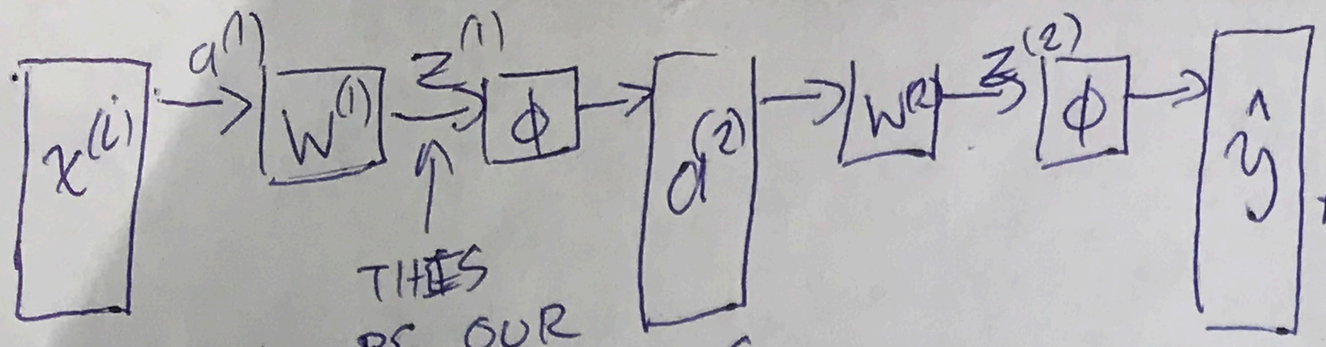
$$\frac{\partial f}{\partial x} = 2(1+\ln(x^3))\frac{\partial}{\partial x}(\cancel{1}^{0} + \log(x^3))$$

$$= 2(1+\ln(x^3))\frac{1}{x^3}\frac{\partial}{\partial x}x^3$$

$$= 2(1+\ln(x^3))\frac{1}{x^3}\,3x^2$$

$$= \frac{6(1+\ln(x^3))}{x}$$

SO WE CAN ✓ USE CHAIN RULE ON OUR NEURAL NETWORK. WE WILL USE "z" AS OUR INTERMEDIATE VARIABLE!!



$$z_i^{(l)} = \sum_{j=1}^{S^{(l-1)}} w_{ij}^{(l)} a_j^{(l)}$$

WE NEED

$$\frac{\partial J}{\partial w_{ij}^{(l)}} = \frac{\partial J}{\partial z_i^{(l)}} \frac{\partial z_i^{(l)}}{\partial w_{ij}^{(l)}}$$

$$= \frac{\partial J}{\partial z_i^{(l)}} \frac{\partial}{\partial w_{ij}^{(l)}} \left(\sum_{q=1}^{S^{(l-1)}} w_{iq}^{(l)} a_q^{(l)} \right)$$

IF $q \neq j$ THEN DERIVATIVE IS $\emptyset$

↑ DUMMY VARIABLE

$$= \frac{\partial J}{\partial z_i^{(l)}} a_j^{(l)}$$

DEFINE AS THE "SENSITIVITY", $v_i^{(l)}$

$$= v_i^{(l)} a_j^{(l)}$$

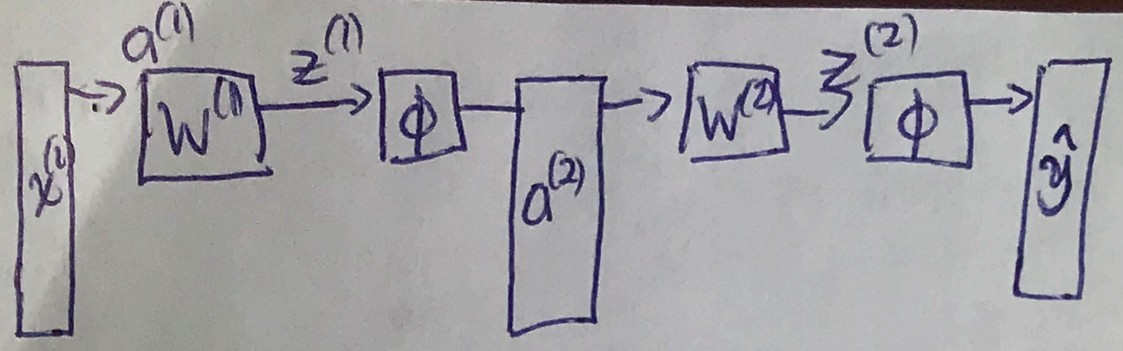
$$W^{(l)} \leftarrow W^{(l)} + \eta \vec{v}^{(l)} \left[\vec{a}^{(l)} \right]^T$$

$$\begin{bmatrix} v_1^{(l)} \\ v_2^{(l)} \\ \vdots \\ v_{S^l}^{(l)} \end{bmatrix} \begin{bmatrix} a_1^{(l)} & a_2^{(l)} & \cdots & a_{S^{(l-1)}} \end{bmatrix} = \begin{bmatrix} a_1 v_1 & a_2 v_1 & \cdots & a_{S^{l-1}} v_1 \\ a_1 v_2 & & & \vdots \\ \vdots & & & \\ a_1 v_{S^{(l)}} & & & a_{S^{(l-1)}} v_{S^{(l)}} \end{bmatrix}$$

$S^l \times S^{(l-1)}$

BUT WE NEED TO ADD UP THE RESPONSE FOR ALL ACTIVATIONS FROM ALL INSTANCES

$$W^{(l)} \leftarrow W^{(l)} + \eta \sum_{R=1}^{M} \left[\vec{v}^{(l)} \right]^{(R)} \left[\vec{a}^{(l)T} \right]^{(R)}$$



$$z_i^{(\ell)} = \sum_{j=1}^{s^{(\ell-1)}} w_{ij}^{(\ell)} a_j^{(\ell)}$$

$$\frac{\partial J}{\partial w_{ij}^{(\ell)}} = \frac{\partial J}{\partial z_i^{(\ell)}} \frac{\partial z_i^{(\ell)}}{\partial w_{ij}^{(\ell)}}$$

$$= \frac{\partial J}{\partial z_i^{(\ell)}} \frac{\partial}{\partial w_{ij}^{(\ell)}} \left(\sum_{q=1}^{s^{(\ell-1)}} w_{iq}^{(\ell)} a_q^{(\ell)} \right)$$

$$= \frac{\partial J}{\partial z_i^{(\ell)}} a_j^{(\ell)}$$

i^{TH} SENSITIVITY, $v_i^{(\ell)}$

$$= v_i^{(\ell)} a_j^{(\ell)}$$

$$W^{(\ell)} \leftarrow W^{(\ell)} + \eta \, \vec{v}^{(\ell)} \left[\vec{a}^{(\ell)}\right]^T$$

$$\begin{bmatrix} v_1^{(\ell)} \\ v_2^{(\ell)} \\ \vdots \\ v_{s^\ell}^{(\ell)} \end{bmatrix} \begin{bmatrix} a_1^{(\ell)} & \cdots & a_{s^{\ell-1}}^{(\ell)} \end{bmatrix} = \begin{bmatrix} a_1 v_1 & a_2 v_1 & \cdots & a_{s^{\ell-1}} v_1 \\ a_1 v_2 & a_2 v_2 & & \\ \vdots & & & \\ a_1 v_{s^{(\ell)}} & \cdots & \cdots & a_{s^{(\ell-1)}} v_{s^{(\ell)}} \end{bmatrix}$$

$s^\ell \times s^{\ell-1}$

BUT WE NEED TO DO THIS FOR ALL ACTIVATIONS

$x^{(1)} \quad x^{(2)} \quad \cdots \quad x^{(M)}$

$$W^{(\ell)} \leftarrow W^{(\ell)} + \eta \sum_{k=1}^{M} \left[\vec{v}^{(\ell)}\right]^{(k)} \left[\left[\vec{a}^{(\ell)}\right]^T\right]^{(k)}$$

$$V^{(\ell)} = \begin{bmatrix} | & & | \\ \left[\vec{v}^{(\ell)}\right]^{(1)} & \cdots & \left[\vec{v}^{(\ell)}\right]^{(M)} \\ | & & | \end{bmatrix}$$

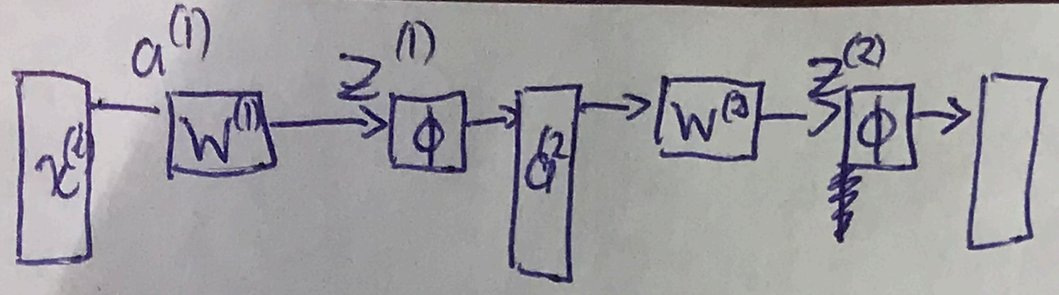
$$W^{(\ell)} \leftarrow W^{(\ell)} + \eta \, V^{(\ell)} \left[A^{(\ell)}\right]^T$$

$(s^\ell \times M)(M \times s^{\ell-1})$

$$A^{(\ell)T} = \begin{bmatrix} - \left[a^{(\ell)}\right]^{(1)} - \\ \vdots \\ - \left[a^{(\ell)}\right]^{(M)} - \end{bmatrix}$$

$$W^{(\ell)} \leftarrow W^{(\ell)} + \eta \, V^{(\ell)} \left[A^{(\ell)}\right]^T$$

HOW DO WE CALCULATE THE SENSITIVITIES?



$$\bar{v}^{(l)} = \frac{\partial J}{\partial \vec{z}^{(l)}} = \frac{\partial J}{\partial \vec{z}^{(l+1)}} \frac{\partial \vec{z}^{(l+1)}}{\partial \vec{z}^{(l)}} = \left[\frac{\partial \vec{z}^{(l+1)}}{\partial z^{(l)}}\right]^T \frac{\partial J}{\partial \vec{z}^{(l+1)}}$$

CHAIN RULE FOR MATRICES OF EQUATIONS IS DEFINED BY JACOBIAN

JACOBIAN D

$$v^{(l)} = D^T v^{(l+1)}$$

THIS IS A RECURRENCE RELATION!

WE CAN BACK PROPAGATE THE SENSITIVITY!

$$D_{ij} = \frac{\partial z_i^{(l+1)}}{\partial z_j^{(l)}} = \begin{bmatrix} \frac{\partial z_1^{(l+1)}}{\partial z_1^{(l)}} & \frac{\partial z_1^{(l+1)}}{\partial z_2^{(l)}} & \cdots & \frac{\partial z_1^{(l+1)}}{\partial z_{s(l)}^{(l)}} \\ \frac{\partial z_2^{(l+1)}}{\partial z_1^{(l)}} & & & \\ \vdots & & & \\ \frac{\partial z_{s(l+1)}^{(l+1)}}{\partial z_1^{(l)}} & & & \frac{\partial z_{s(l+1)}^{(l+1)}}{\partial z_s^{(l)}} \end{bmatrix}$$

$$\frac{\partial z_i^{(l+1)}}{\partial z_j^{(l)}} = \frac{\partial}{\partial z_j^{(l)}} \left(\sum_{q=1}^{s^{(l)}} w_{iq}^{(l+1)} a_q^{(l+1)} \right)$$

$$= \frac{\partial}{\partial z_j^{(l)}} \sum_{q=1}^{s^{(l)}} w_{iq}^{(l+1)} \phi(z_q^{(l)}) \qquad q = j \text{ for nonzero!}$$

$$= w_{ij}^{(l+1)} \phi_j' \qquad \text{if } \phi \text{ is sigmoid } \phi_j' = a_j(1-a_j)$$

FOR SIG. $D_{ij} = w_{ij}^{(l+1)} a_j^{(l+1)} (1 - a_j^{(l+1)})$

$$D = W^{(l+1)} \dot{\Phi}^{(l+1)}$$

$$\dot{\Phi}' = DIAG(\dot{\phi}') = \begin{bmatrix} \dot{\phi}_1^{(l+1)} & & \\ & \dot{\phi}_2^{(l+1)} & \\ & & \ddots \\ & & & \dot{\phi}_{s^{(l+1)}}^{(l+1)} \end{bmatrix}$$

$$v^{(l)} = D^T v^{(l+1)}$$

$$v^{(l)} = \left[W^{(l+1)} \dot{\Phi}^{(l+1)}\right]^T v^{(l+1)} = \dot{\Phi}^{(l+1)} \left[W^{(l+1)}\right]^T \vec{v}^{(l+1)}$$

$$= \dot{\phi}^{(l+1)} \circledast \left[W^{(l+1)}\right]^T \vec{v}^{(l+1)} = \begin{bmatrix} \dot{\phi} \\ \vdots \\ \dot{\phi} \end{bmatrix} = \begin{bmatrix} wv \end{bmatrix} = \begin{bmatrix} 1 \end{bmatrix}$$

$$V^{(l)} = \begin{bmatrix} | & | & & | \\ [\vec{v}^{(l)}]^{(1)} & [v^{(l)}]^{(2)} & \cdots & [v^{(l)}]^{(M)} \\ | & | & & | \end{bmatrix}$$

$$= \begin{bmatrix} | & | & & | \\ [\vec{\phi}^{(l+1)}]^{(1)} \circledast [W^{(l+1)}][\vec{v}^{(l+1)}]^{(1)} & \cdots & [\vec{\phi}^{(l+1)}]^{(M)} \circledast [W^{(l+1)}][v^{(l+1)}]^{(M)} \\ | & & | \end{bmatrix}$$

$$= \underset{(S^l \times M)}{\dot{\Phi}^{(l+1)}} \circledast \underbrace{[W^{(l+1)}]}_{(S^l \times S^{(l+1)})}{}^{T} \underbrace{V^{(l+1)}}_{(S^{l+1} \times M)}$$

$$\dot{\phi} = \vec{a}(1-\vec{a})$$

$$\dot{\Phi}^{(l+1)} = A^{(l+1)}(1-A^{(l+1)})$$

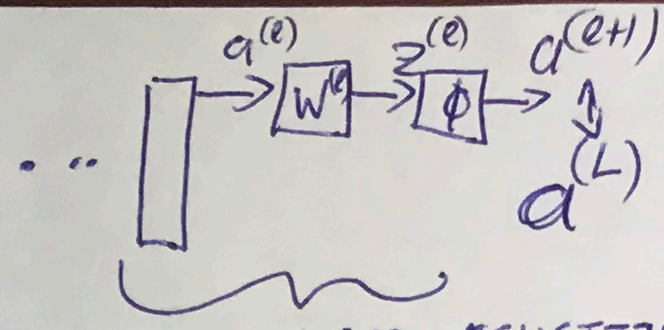
$$V^{(l)} = A^{(l+1)}(1-A^{(l+1)}) \circledast [W^{(l+1)}]^{T} V^{(l+1)}$$

FAST CALCULATION OF RECURRENCE RELATION!

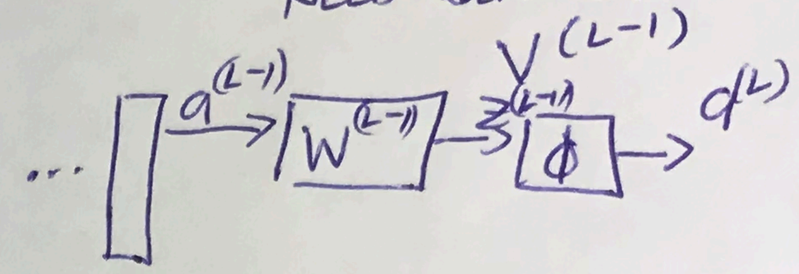
NEED $V^{(L)}$ at THE LAST LAYER

- $J(w) = \sum_{k=1}^{M} \left| \vec{y}^{(k)} - [a^{(L)}]^{(k)} \right|^2$

$$= \sum_{k=1}^{M} \left| y^{(k)} - \phi([z^{(L-1)}]^{(k)}) \right|^2$$



NEED SENSTIVITY FOR



$$\left[\nabla_i^{(L-1)}\right]^{(k)} = \left[\frac{\partial J}{\partial z_i^{(L-1)}}\right]^{(k)}$$

~~$= \frac{\partial}{\partial z_i^{(L-1)}} \sum_{k=1}^{M} (\vec{y}^{(k)} - \phi([\vec{z}^{(L-1)}]^{(k)}))^2$~~

$$= \frac{\partial}{\partial z_i^{(L-1)}} \left(y^{(k)} - \phi([z^{(L-1)}]^{(k)})\right)^2$$

$$= \frac{\partial}{\partial z_i^{(L-1)}} \left(y_i^{(k)} - \phi([z_i^{(L-1)}]^{(k)})\right)^2$$

$$= (y_i^{(k)} - \phi_i(\cdot)) \, 2 \left(-\frac{\partial}{\partial z_i^{(L-1)}} \phi(z_i^{(L-1)})\right)$$

$$= -2 (y_i^{(k)} - \phi_i(\cdot)^{(k)}) [a_i^{(L)}]^{(k)} (1 - [a_i^{(L)}]^{k})$$

~~$\vec{V}^{(L-1)} = -2(\vec{y} - \vec{\phi}) * \vec{a}^{(L)} * (1 - \vec{a}^{(L)})$~~

~~$V^{(L-1)} = -2 Y - \phi) * a^{(L)} * (1 - a^{(L)})$~~

$$= -2 (y_i^{(k)} - [a_i^{(L)}]^{(k)}) [a_i^{(L)}]^{(k)} (1 - [a_i^{(L)}]^{(k)})$$

$$\left[\vec{V}^{(L-1)}\right]^{(k)} = -2 (\vec{y}^{(k)} - [\vec{a}^{(L)}]^{(k)}) * [\vec{a}^{(L)}]^{(k)} * (1 - [\vec{a}^{(L)}]^{(k)})$$

$$V^{(L-1)} = -2 (Y - A^{(L)}) * A^{(L)} * (1 - A^{(L)})$$

$$V^{(L-1)} = \begin{bmatrix} 1 & & 1 \\ [\vec{V}^{L-1}]^{(1)} & \cdots & [\vec{V}^{(L-1)}]^{(M)} \\ 1 & & 1 \end{bmatrix}$$

↑ SOLVE 1ST

CHAIN RULE!

SUMMARY

① PERFORM FEED FORWARD

$$X^T = A^{(1)} \qquad Z^{(1)} = W^{(1)} A^{(1)}$$

$$Z^{(\ell)} = W^{(\ell)} A^{(\ell)}$$

$$\phi(Z^{(\ell)}) = A^{(\ell+1)}$$

② GET FINAL LAYER SENSITIVITY

$$V^{(L-1)} = -2(Y - A^{(L)}) * A^{(L)} * (1 - A^{(L)})$$

③ BACK PROP. ALL SENS.

$$V^{(\ell)} = A^{(\ell+1)} \circledast (1 - A^{(\ell+1)}) * [W^{(\ell+1)}]^T V^{(\ell+1)}$$

④ UPDATE ALL WEIGHT MATRICES

$$W^{(\ell)} \Leftarrow W^{(\ell)} + \eta V^{(\ell)} [A^{(1)}]^T$$

- Self Test: We assumed that no bias term was in each layer. Can we update our notation to include bias vectors for each layer?