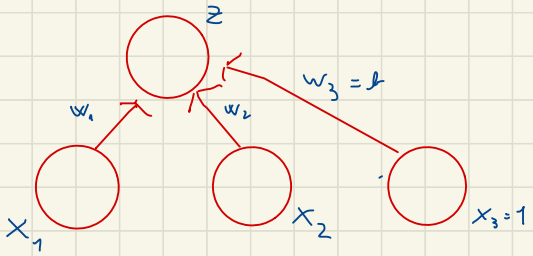
# Backpropagation

# GRADIENT DESCENT & LINEAR REGRESSION




Compute $z$ in forward propagation:

$$z^{(p)} = x_1^{(p)} w_1 + x_2^{(p)} w_2 + b = \vec{w} \cdot \vec{x}^{(p)}$$

Compute $J(w) = \sum_{p \in T_n} \left( t^{(p)} - z^{(p)} \right)^2$

Let's compute the gradient for $w_1$

$$\frac{\partial J}{\partial w_1} = \frac{\partial}{\partial w_1} \frac{1}{2 N_{Tn}} \sum_{p \in T_n} \left( t^{(p)} - z^{(p)} \right)^2 \quad \text{Sum rule}$$

$$= \frac{1}{2 N_{Tn}} \sum_{p \in T_n} \frac{\partial}{\partial w_1} \left( t^{(p)} - z^{(p)} \right)^2 \quad \text{Power rule}$$

$$= \frac{1}{2 N_{Tn}} \sum_{p \in T_n} 2 \left( t^{(p)} - z^{(p)} \right) \frac{\partial}{\partial w_1} \left( t^{(p)} - z^{(p)} \right)$$

$\frac{\partial}{\partial w_1} t^{(p)} = 0$ $\quad \frac{\partial}{\partial w_1} z^{(p)} = \frac{\partial}{\partial w_1} \left( x_1^{(p)} w_1 + x_2^{(p)} w_2 + b \right) = x_1^{(p)}$

$$= -\frac{1}{N_{Tn}} \sum_{p \in T_n} \left( t^{(p)} - z^{(p)} \right) x_1^{(p)}$$

Similar derivation for $w_2$ and $b$.

$$\nabla_{\vec{w}} J = \left[ \frac{\partial J}{\partial w_1}, \frac{\partial J}{\partial w_2}, \frac{\partial J}{\partial w_3} \right]$$

MATRIX NOTATION

$$\nabla_{\vec{w}} J = \frac{1}{N_{Tn}} \sum_{p \in T_n} \left( t^{(p)} - z^{(p)} \right) \vec{x}^T$$

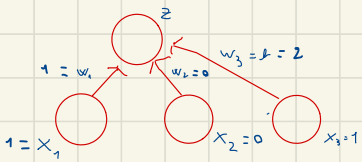
For a single example:

$$\frac{dJ}{dz} \frac{dz}{d\vec{w}} \rightarrow \nabla_{\vec{w}} z$$

For gradient row notation

Let's compute a numerical example.

$T_n = \{ ([1, 0, 2]^T, 2) \}$ ← single example. EXERCISE: TRY WITH 2!




FORWARD: $z = 1 \cdot 1 + 0 \cdot 0 + 2 \cdot 1 = 3$

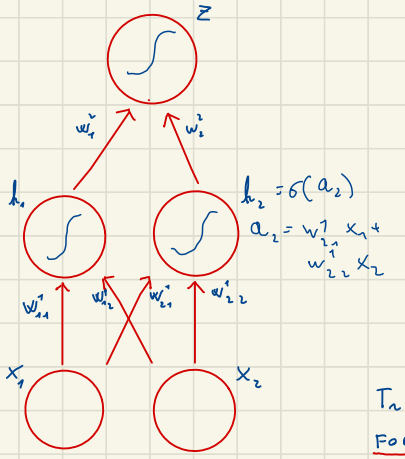
$$\frac{\partial J}{\partial w_1} = -(t - z) x_1 = (2 - 3) \cdot 1 = +1$$

$$\frac{\partial J}{\partial w_2} = -(t - z) x_2 = (2 - 3) \cdot 0 = 0 \qquad \frac{\partial J}{\partial w_3} = (2 - 3) \cdot 1 = +1$$

MATRIX NOTATION $\quad \nabla_{\vec{w}} J = -(t - z) \vec{x} = (2 - 3) [1, 0, 1] = [+1, 0, +1]$ ■

We can update $\vec{w} \leftarrow \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} - \underbrace{\epsilon}_{0.5} \begin{bmatrix} +1 \\ 0 \\ +1 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0 \\ 1.5 \end{bmatrix}$ NOTE: different learning rates would require more iterations!

# BACK PROPAGATION




RECALL:

$\sigma(NET) = \frac{1}{1+e^{-NET}}$ $\quad \sigma'(NET) = \frac{d\,\sigma(NET)}{d\,NET} = \sigma(NET)(1-\sigma(NET))$

CHAIN RULE:

$$\frac{d\,f(y(x))}{dx} = \frac{d\,f(y(x))}{d\,y(x)} \frac{d\,y(x)}{dx}$$

$$\vec{a} = \begin{bmatrix} w_{11}^1 & w_{12}^1 \\ w_{21}^1 & w_{22}^1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$T_n = \{(x^{(1)}, t^{(1)}) \dots (x^{(N_{Tn})}, t^{(N_{Tn})})\}$

FORWARD PROP: COMPUTES $z$ and the hidden representation $\vec{h}$

Let's consider MSE loss: $J = \frac{1}{2} \frac{1}{N_{Tn}} \sum_{p \in T_n} (t^{(p)} - z^{(p)})^2$

Let's compute the gradient w.r.t. the output weights $\vec{w}_1^2$. $\vec{w}_2^2$ is analogous

$\frac{\partial J}{\partial w_1^2} = \frac{\partial}{\partial w_1^2} \frac{1}{2 N_{Tn}} \sum_{p \in T_n} (t^{(p)} - z^{(p)})^2$ derivative is linear. Derivative of sum is the sum of the derivatives.

$= \frac{1}{2 N_{Tn}} \sum_{p \in T_n} \frac{\partial}{\partial w_1^2} (t^{(p)} - z^{(p)})^2$ power rule: $\frac{d}{dx} f(x)^m = m[f(x)]^{m-1} f'(x)$

$= \frac{1}{2 N_{Tn}} \sum_{p \in T_n} 2 (t^{(p)} - z^{(p)}) \frac{\partial}{\partial w_1^2} (t^{(p)} - z^{(p)})$ $z^{(p)} = \sigma(w_1^2 h_1^{(p)} + w_2^2 h_2^{(p)}) = \sigma(\vec{w}^2 \cdot \vec{h}^{(p)})$

$= \frac{1}{N_{Tn}} \sum_{p \in T_n} (t^{(p)} - z^{(p)}) \frac{\partial}{\partial w_1^2} (t^{(p)} - \sigma(\vec{w}^2 \cdot \vec{h}^{(p)}))$ sum rule. $\frac{\partial}{\partial w_1^2} t^{(p)} = 0$ chain rule: $\frac{\partial}{\partial w_1^2} \sigma(\vec{w}^2 \cdot \vec{h}^{(p)}) = \sigma'(\vec{w}^2 \cdot \vec{h}^{(p)}) \cdot h_1^{(p)}$ (SIGMOID)

$= -\frac{1}{N_{Tn}} \sum_{p \in T_n} (t^{(p)} - z^{(p)}) \sigma'(\vec{w}^2 \cdot \vec{h}^{(p)}) h_1^{(p)}$ $\nabla_{z^{(p)}} J$ BACKWARD PASS $f'(a^{(p)})$ FORWARD PASS

For sigmoid: $-\frac{1}{N_{Tn}} \sum_{p \in T_n} (t^{(p)} - z^{(p)}) \sigma(\vec{w}^2 \cdot \vec{h}^{(p)}) (1 - \sigma(\vec{w}^2 \cdot \vec{h}^{(p)})) h_1^{(p)}$

$\frac{\partial \sigma(\vec{w}^2 \cdot \vec{h}^{(p)})}{\partial \vec{w}^2 \cdot \vec{h}^{(p)}} =$

$\sigma(\vec{w}^2 \cdot \vec{h}^{(p)})(1 - \sigma(\vec{w}^2 \cdot \vec{h}^{(p)}))$

$\begin{bmatrix} w_{1,1}^1 & w_{1,2}^1 \\ w_{2,1}^1 & w_{2,2}^1 \end{bmatrix}$

Let's now compute the gradient for the hidden units. $w_{1,1}^1$

$\frac{\partial J}{\partial w_{1,1}^1} = \frac{\partial}{\partial w_{1,1}^1} \frac{1}{2 N_{Tn}} \sum_{p \in T_n} (t^{(p)} - z^{(p)})^2$ Sum Rule

$= \frac{1}{2 N_{Tn}} \sum_{p \in T_n} \frac{\partial}{\partial w_{1,1}^1} (t^{(p)} - z^{(p)})^2$ Power Rule and $\frac{\partial}{\partial w_{1,1}^1} t^{(p)} = 0$ and Sum Rule

$= \frac{1}{2 N_{Tn}} \sum_{p \in T_n} 2 (t^{(p)} - z^{(p)}) \frac{\partial}{\partial w_{1,1}^1} (-z^{(p)})$ $z^{(p)} = \sigma(\vec{w}^2 \cdot \vec{h}^{(p)})$ and Chain Rule

$= \frac{1}{2 N_{Tn}} \sum_{p \in T_n} 2 (t^{(p)} - z^{(p)}) \sigma'(\vec{w}^2 \cdot \vec{h}^{(p)}) \frac{\partial}{\partial w_{1,1}^1} \vec{w}^2 \cdot \vec{h}^{(p)}$ $\vec{w}^2 \cdot \vec{h}^{(p)} = w_1^2 h_1^{(p)} + w_2^2 h_2^{(p)}$

$= \frac{1}{2 N_{Tn}} \sum_{p \in T_n} 2 (t^{(p)} - z^{(p)}) \sigma'(\vec{w}^2 \cdot \vec{h}^{(p)}) \frac{\partial}{\partial w_{1,1}^1} w_1^2 h_1^{(p)} + \underbrace{\frac{\partial}{\partial w_{1,1}^1} w_2^2 h_2^{(p)}}_{=0}$ $h_1^{(p)} = \sigma(\vec{w}^1 \cdot x^{(p)})$

$\frac{dJ}{dz} \frac{dz}{da^2} \frac{da_2}{d\vec{w}^2}$ $\quad \vec{h}^T$

$\nabla_{\vec{w}^2} a_2 = [h_1, h_2]$

$$= \frac{1}{N_{TR}} \sum_{p \in T_R} (t^{(p)} - z^{(p)}) \sigma'(\vec{w}^2 \vec{h}^{(p)}) \; w_1^2 \; \frac{\partial}{\partial w_{11}^1} \sigma\left(\sum_{k=1}^{2} w_{1k}^1 x_k + w_{12}^1 x_2\right)$$ CHAIN RULE

$$= \frac{1}{N_{TR}} \sum_{p \in T_R} (t^{(p)} - z^{(p)}) \sigma'(\vec{w}^2 \vec{h}^{(p)}) \; w_1^2 \; \sigma'(\underbrace{\vec{w}_1^1}_{\begin{bmatrix} w_{11}^1 \\ w_{12}^1 \end{bmatrix}} \vec{x}^{(p)}) \; \frac{\partial}{\partial w_{11}^1} (\underbrace{w_{11}^1 x_1^{(p)}}_{= x_1^{(p)}} + \underbrace{w_{12}^1 x_2^{(p)}}_{= 0 \text{ constant}})$$

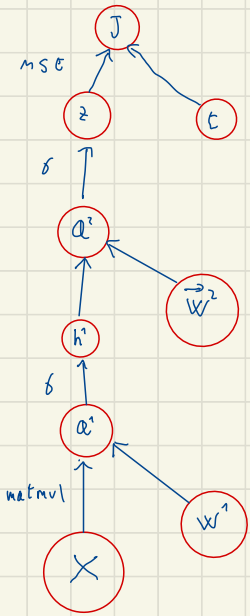
MATRIX NOTATION

on a single example for simplicity

$$\frac{dJ}{dw^2} = \frac{dJ}{dz} \frac{dz}{da^2} \frac{da_2}{dw^2} =$$

$$= -\left[(t - z)\right] \left[\sigma(a^2)(1 - \sigma(a^2))\right] \nabla_{\vec{w}^2} a^2$$

$$\nabla_{\vec{w}^2} a^2 = [h_1, h_2]$$




$$\frac{dJ}{dw^1} = \underbrace{\frac{dJ}{dz}}_{\mathbb{R}} \underbrace{\frac{dz}{da^2}}_{\mathbb{R}} \underbrace{\frac{da^2}{dh}}_{\mathbb{R}^{1\times 2}} \underbrace{\frac{dh}{da^1}}_{\mathbb{R}^{2\times 2}} \underbrace{\frac{da^1}{dw^1}}_{\mathbb{R}^{2\times(2\times 2)}} \rightarrow \mathbb{R}^{2\times(2\times 2)}$$

$$= -\left[(t - z)\right] \left[\sigma(a^2)(1 - \sigma(a^2))\right] \underbrace{\begin{bmatrix} w_1^2 & w_2^2 \end{bmatrix}}_{= \begin{bmatrix} \frac{\partial a^2}{\partial h_1} & \frac{\partial a^2}{\partial h_2} \end{bmatrix}} \begin{bmatrix} \frac{dh_1}{da_1^1} & \frac{dh_1}{da_2^1} \\ \frac{dh_2}{da_1^1} & \frac{dh_2}{da_2^1} \end{bmatrix} \begin{bmatrix} \frac{\partial a_1^1}{\partial w^1} \\ \frac{\partial a_2^1}{\partial w^1} \end{bmatrix}$$

$$a_1^1 = w_{11}^1 x_1 + w_{12}^1 x_2$$

$$\frac{\partial a_1^1}{\partial w_{1,:}^1} = x_:$$

$$\frac{\partial a_1^1}{\partial w_{1,:}^1} = \vec{x}^T \in \mathbb{R}^{1\times 1\times 2}$$

$$\frac{\partial a_1^1}{\partial w_{k\neq 1,:}^1} = \vec{0}^T \in \mathbb{R}^{1\times 1\times 2}$$

$$\begin{bmatrix} \frac{dh_1}{da_1^1} & \frac{dh_1}{da_2^1} \\ \frac{dh_2}{da_1^1} & \frac{dh_2}{da_2^1} \end{bmatrix} = \begin{bmatrix} \sigma(a_1^1)(1 - \sigma(a_1^1)) & 0 \\ 0 & \sigma(a_2^1)(1 - \sigma(a_2^1)) \end{bmatrix}$$

$$\begin{bmatrix} x_1 & x_2 \\ 0 & 0 \end{bmatrix} \quad \frac{\partial a_1^1}{\partial \vec{w}_1^1}, \; \frac{\partial a_2^1}{\partial \vec{w}_2^1}$$

$$\begin{bmatrix} 0 & 0 \\ x_1 & x_2 \end{bmatrix}$$

JACOBIAN IS DIAGONAL SINCE ACT. FUNCTION IS ELEMENT-WISE

ELEMENT-WISE MULT. WITH DIAGONAL

$$\odot \left[\sigma(\,) w_1^2, \; \sigma(\,) w_2^2\right]$$

VERIFY CORRECTNESS OF MATRIX FORMULATION ( DIFFICULT : PRODUCT WITH TENSORS)

UNF. OF TENSOR

$$\frac{da^1}{dw^1}\begin{bmatrix} x_1 & x_2 & 0 & 0 \\ 0 & 0 & x_1 & x_2 \end{bmatrix}$$

$$(t-z)\,(\sigma(a^2)(1-\sigma(a^2)))\begin{bmatrix} w^2_1\,\sigma(a^1_1)(1-\sigma(a^1_1)) & , & w^2_2\,\sigma(a^1_2)(1-\sigma(a^1_2)) \end{bmatrix} \cdot \frac{da^1}{dw^1}$$

$$\in \mathbb{R}^{1\times 2\times 2}$$

$$= \begin{bmatrix} w^2_1\,\sigma(\dots)\,x_1 & , & w^2_1\,\sigma(\dots)\,x_2 & , & w^2_2\,\sigma(\dots)\,x_1 & , & w^2_2\,\sigma(\dots)\,x_2 \end{bmatrix}$$

$$= \begin{bmatrix} w^2_1\,\sigma(\dots)\,x_1 & w^2_1\,\sigma(\dots)\,x_2 \\ w^2_2\,\sigma(\dots)\,x_1 & w^2_2\,\sigma(\dots)\,x_2 \end{bmatrix}$$

OR EQUIVALENTLY

$$\left( (t-z)\,(\sigma(a^2)(1-\sigma(a^2)))\,\underbrace{\begin{bmatrix} w^2_1 & w^2_2 \end{bmatrix}}_{w^{2T}} \odot \begin{bmatrix} \sigma(a^1_1)(1-\sigma(a^1_1)) & , & \sigma(a^1_2)(1-\sigma(a^1_2)) \end{bmatrix} \right)^T x^T$$

EXTENTION TO MULTIPLE EXAMPLES IN MATRIX NOTATION

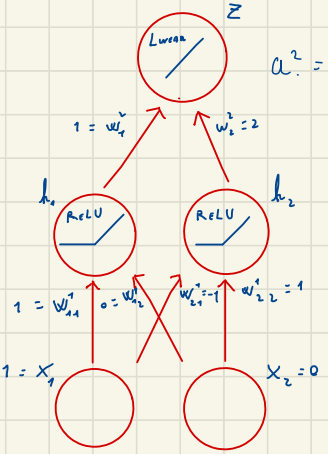
$$J = \frac{1}{n}\begin{bmatrix} J^{(1)} & \dots & J^{(n)} \end{bmatrix}\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

what we computed before for a single example

## EXERCISE

ON A SINGLE EXAMPLE!

Let's consider an istantiation of our NN, and let's compute the gradients numerically ✓ and update the weights.




$a^2 = w^2 \vec{h}$ $\quad z = i(a^2) = a^2$

DERIVATIVE OF LINEAR

$l(x) = 1$

DER. of ReLU

$$\text{ReLU}'(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Let's assume $c = 2$

$N_{Tr} = 1$

Let's start with the FORWARD propagation.

$h_1 = \max(0, 1 \cdot 1 + 0 \cdot 0) = \max(0, 1) = 1$

$h_2 = \max(0, 1 \cdot -1 + 0 \cdot 1) = \max(0, -1) = 0$

$z = 1 \cdot 1 + 2 \cdot 0 = 1$

$J = (t - z)^2 = 1$

$$\frac{\partial J}{\partial w_1^2} = \frac{1}{N_{Tr}} \sum_{i \in Tr} -(t^{(i)} - z^{(i)}) \underbrace{g'(\vec{w}^2 \cdot \vec{h}^{(i)})}_{\text{linear}} h_1^{(i)} = -(t - z) \cdot 1 \cdot h_1 = (z - t) \cdot 1 \cdot 1 = -1$$

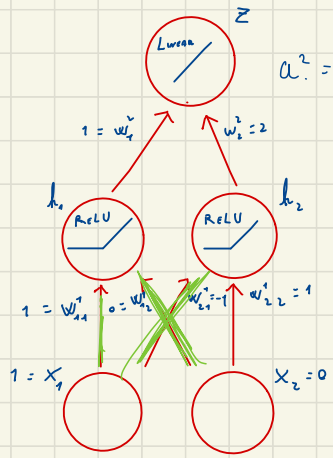
($a^2$ over $\vec{w}^2 \cdot \vec{h}^{(i)}$)

$$\frac{\partial J}{\partial w_2^2} = (z - t) \cdot 1 \cdot 0 = 0$$

**EXERCISE**

Compute the partial derivatives w.r.t. the weights of the first layer

... compute them also in matrix notation




$\frac{d}{z}\frac{1}{2}(t-z)^2 = -(t-z)$

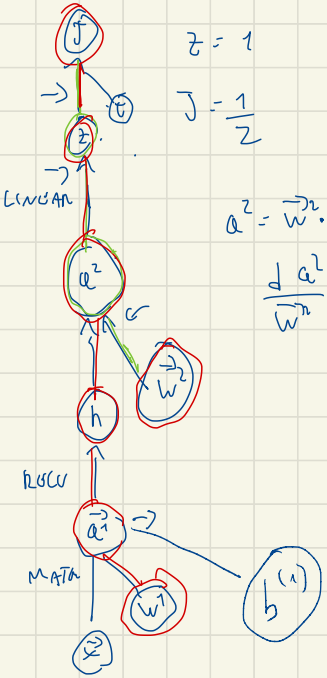
$t = 2$

$\frac{d\,MSE}{z} = \frac{1}{2}2(t-z)(-1)$

$J = MSE = \frac{1}{2}(1-2)^2$

$x = \begin{bmatrix}1\\0\end{bmatrix}$ $h = \begin{bmatrix}1\\0\end{bmatrix}$

Let's start with the FORWARD




$z = 1$

$J = \frac{1}{2}$

$a^2 = \vec{w}^2 \cdot \vec{h}$

$\frac{da^2}{\vec{w}^2} = \vec{h}$

$\frac{\partial J}{\partial \vec{w}^2} = \frac{dJ}{dz}\frac{dz}{da^2}\frac{da^2}{d\vec{w}^2}$

$-1 \quad 1 \quad h = \begin{bmatrix}1\\0\end{bmatrix}^T$

$= \begin{bmatrix}-1\\0\end{bmatrix}$

$\vec{w}^2 = \vec{w}^2 - \alpha\begin{bmatrix}-1\\0\end{bmatrix}$

$\begin{bmatrix}1\\2\end{bmatrix} + \begin{bmatrix}1\\0\end{bmatrix} = \begin{bmatrix}2\\2\end{bmatrix}$

$\frac{\partial J}{\partial w^1} = \frac{dJ}{dz}\frac{dz}{da^2}\frac{da^2}{dh}\frac{dh}{da^1}\frac{da^1}{dw^1}$

$\frac{dh}{da^1} = \begin{bmatrix}\frac{dh_1}{da_1} & 0\\ 0 & \frac{dh_2}{da_2}\end{bmatrix}$

$\vec{w}^2\begin{bmatrix}1 & 2\end{bmatrix}$

$\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix} = 0\begin{bmatrix}1\\0\end{bmatrix} = 2 \cdot 0 = \begin{bmatrix}2\\0\end{bmatrix}$

$\frac{da^1}{dw^1} = R^{2\times 2\times 1}$

Trick $\nearrow x^T$

$\begin{bmatrix}-1\\0\end{bmatrix}\begin{bmatrix}1 & 0\end{bmatrix}$

$-1 \cdot 1 \cdot \begin{bmatrix}1 & 2\end{bmatrix} \circ \begin{bmatrix}1 & 0\end{bmatrix} = \begin{bmatrix}-1 & 0\end{bmatrix}$

$\frac{dJ}{dw^1} = \begin{bmatrix}-1 & 0\\ 0 & 0\end{bmatrix}$