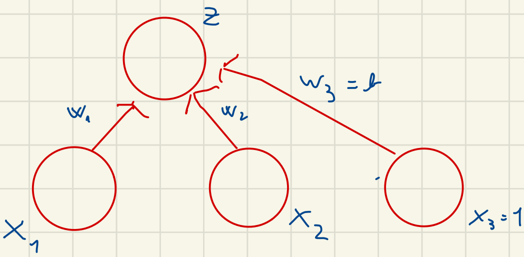
GRADIENT DESCENT & LINEAR REGRESSION



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 $J(w) = \sum_{p \in T_r} (t^{(p)} - z^{(p)})^2$

Let's compute the gradient for w_1

$$\frac{\partial J}{\partial w_1} = \frac{\partial}{\partial w_1} \frac{1}{2 N_{Tr}} \sum_{p \in T_r} (t^{(p)} - z^{(p)})^2$$ Sum rule

$$= \frac{1}{2 N_{Tr}} \sum_{p \in T_r} \frac{\partial}{\partial w_1} (t^{(p)} - z^{(p)})^2$$ Power Rule

$$= \frac{1}{2 N_{Tr}} \sum_{p \in T_r} 2 (t^{(p)} - z^{(p)}) \frac{\partial}{\partial w_1} (t^{(p)} - z^{(p)})$$

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

$$= \boxed{-} \frac{1}{\cancel{2} N_{Tr}} \sum_{p \in T_r} \cancel{2} (t^{(p)} - z^{(p)}) x_1^{(p)}$$

Similar derivation for w_2 and b.

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

FOR A SINGLE EXAMPLE: $\frac{\partial J}{\partial z} \frac{\partial z}{\partial \vec{w}} \to \nabla_{\vec{w}} z$

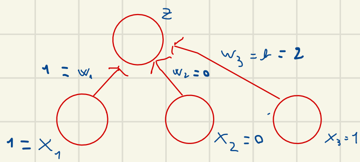
For gradient row notation

MATRIX NOTATION

$$\nabla_{\vec{w}} J = \frac{1}{N_{Tr}} \sum_{p \in T_r} (t^{(p)} - z^{(p)}) \vec{x}^T$$

Let's compute a numerical example.

$T_r = \left\{ \left(\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, 2 \right) \right\}$ ← SINGLE EXAMPLE. EXERCIZE: TRY WITH 2!



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

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

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

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

We can update $\vec{w} \leftarrow \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} - \underset{0.5}{\epsilon} \begin{bmatrix} +1 \\ 0 \\ +1 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0 \\ 1.5 \end{bmatrix}$

NOTE: different learning rates would require more iterations!