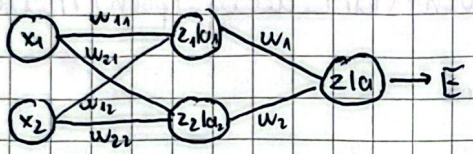
Section 7: Weight Initialization

- If we initialize all weighs to 0, knowing that the updates to the weights is dependant on previous weights. And since ~~none~~ all of our weights is 0, no learning will occur

Ex: Our NN in Section 1, $\frac{\partial \mathcal{E}}{\partial w_{11}} = \left\{ \sum_{j=1}^{2} \left[\frac{\partial \mathcal{E}}{\partial \sigma(y_j)} \cdot \frac{\partial \sigma(y_j)}{\partial y_j} \cdot \frac{\partial y_j}{\partial (\sigma(z_1))} \right] \right\} \cdot \frac{\partial (\sigma(z_1))}{\partial z_1} \cdot \frac{\partial z_1}{\partial w_{11}}$

we have $\frac{\partial z_1}{\partial w_{11}} = \frac{\partial y_1}{\partial \sigma(z_1)} = w_{21} = 0$, $\frac{\partial y_2}{\partial \sigma(z_1)} = w_{22} = 0$

- Random initialization: maybe vanish and explode
- Random initialization from Normal Distribution - 1st way
- If we initialize all weights to the same value, for example



$z_1 = x_1 w_{11} + x_2 w_{12}$
$z_2 = x_1 w_{21} + x_2 w_{22}$
$a_1 = \sigma(z_1)$, $a_2 = \sigma(z_2)$
$z_1 = a_1 w_1 + a_2 w_2$, $a = \sigma(z)$

Weight update for w_1: $\frac{dE}{dw_1} = \frac{dE}{da} \cdot \frac{da}{dz} \cdot \frac{dz}{dw_1}$

Weight update for w_2: $\frac{dE}{dw_2} = \frac{dE}{da} \cdot \frac{da}{dz} \cdot \frac{dz}{dw_2}$

$\frac{dz}{dw_1} = a_1 = \sigma(z_1) = \sigma(z_2) = a_2 = \frac{dz}{dw_2}$

Both weight will get updated with the same value and they will be equal. This is called

Symmetry Breaking Problem

- Xavier Initialization: - 2nd Way
  - The output of a linear layer: $Y = W_1 X_1 + W_2 X_2 + \dots + W_n X_n$
  - Assume X_i and W_i are all independent and indentically distributed

$$\begin{aligned} Var(W_i X_i) &= E[X_i]^2 Var(W_i) + E[W_i]^2 Var(X_i) + Var(W_i) Var(X_i) \\ &= Var(W_i) Var(X_i) \end{aligned}$$

Then, $Var(Y) = Var(\sum_i W_i X_i) = n\, Var(W_i)\, Var(X_i)$
We want $Var(Y) = Var(X_i)$, then $Var(W_i) = \frac{1}{n} = \frac{1}{n_{in}}$

Go through same steps for backpropagating, we also have: $Var(W_i) = \frac{1}{n_{out}}$
Compromising: $Var(W_i) = \frac{2}{n_{in} + n_{out}} \Rightarrow std = \sqrt{Var} = \sqrt{\frac{2}{n_{in} + n_{out}}}$

- He Norm Initialization:

A ReLU is zero for half of its input (negative inputs), so we need to double the size of the weight variance to keep the signals variance constant. Therefore, we multiply our previous equation of $\frac{1}{n_{in}}$ by 2.

$$Var(W) = \frac{2}{n_{in}} \qquad std = \sqrt{Var} = \sqrt{\frac{2}{n_{in}}}$$

So,

+ When the layers have ReLU activation → He initialization
+ When the layers have Sigmoid activation → Xavier initialization.