How to initialize weights

- **Don't :**
  - Init to all zero, because gradient will not get updated.

Consider this 2 layers network:

$$x \rightarrow z_1 = W_1 x \rightarrow z_1^+ = \max(z_1, 0) \rightarrow z_2 = W_2 z_1^+ \rightarrow y = z_2$$

If we initialize $W_1 = 0$, then:

$$\Rightarrow \begin{cases} z_1 = 0, \text{ hence } z_1^+ = 0 \\ z_2 = 0 \end{cases}$$

Now lets try to calculate the gradient of W_2:

$$\nabla_{W_2} L_2 = \nabla_{W_2} z_2 \cdot \nabla_{z_2} L_2$$
$$= z_1^+ \cdot -(y - z_2)$$
$$= 0 \cdot -(y - 0)$$
$$= 0$$

$\Rightarrow$ No learning can be done!
(since gradient = 0)

Conclusion: If we have more than 1 layer, and we init weight to zero, no learning can be done.

Solution: Init weight to random values, following either normal distribution or Uniform distribution