Weight initialization: why it matters

$z = Wx + b$, $\quad X \sim P(0, \sigma_x^2)$, X is n samples × k features

$E[z] = E[Wx + b]$, assume X are $\sim N(0,1)$

b = 0 for simplicity

for a single neuron i:

$$E[z_i] = E[w_i x + \cancel{b_i}] = E\left[\sum_{j=1}^{k} w_{ij} x_j\right]$$

$$= E\left[\sum^{k} w_{ij}\right] E\left[x_j\right]$$

← assuming w and x are independent

$$= \qquad 0 \qquad = 0$$

$$\sigma_{z_i} = E[z_i^2] - \cancel{(E[z])^2}$$

$$= E\left[\left(\sum^{k} w_{ij} x_j\right)^2\right] = \sum^{k} E[w_{ij}^2] \underbrace{E[x_j^2]}_{\sigma_x = 1}$$

$$= k \sigma_w^2 \sigma_x^2 \quad //$$

↑ at each layer, variance increases by # of inputs