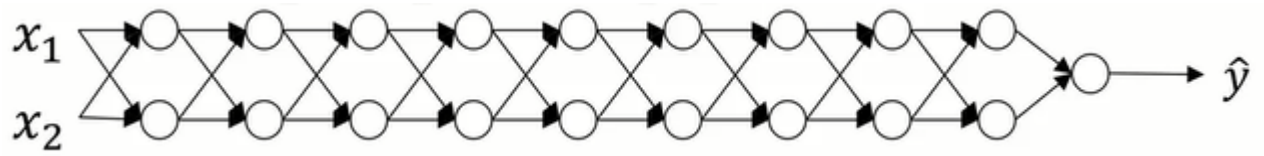
Vanishing/exploding gradients

Example



$g(z) = z,\ b^{[l]} = 0$

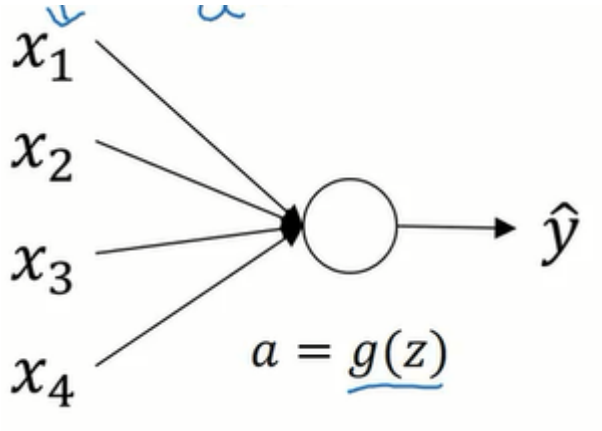
$y = W^{[L]}W^{[L-1]}\dots W^{[2]}W^{[1]}x$

$W^{[l]} = \begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix}$

$y = W^{[L]}\begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix}^{(L-1)}x = W^{[L]}\begin{bmatrix} a^{(L-1)} & 0 \\ 0 & a^{(L-1)} \end{bmatrix}x$

$\begin{cases} exploding\ if\ a > 1 \\ vanishing\ if\ a < 1 \end{cases}$

Weight initialization for deep networks

- 

  - $z = w_1x_1 + w_2x_2 + \dots + w_nx_n$
    - $want\ z\ not\ to\ explode,\ larger\ n \rightarrow smaller\ w_i$
  - $Var(w_i) = \begin{cases} \frac{2}{n},\ when\ g(z)\ is\ ReLU \\ \frac{1}{n},\ when\ g(z)\ is\ tanh \end{cases}$
  - $W^{[l]} = \begin{cases} np.random.randn(shape) * np.sqrt(\frac{2}{n^{[l-1]}}),\ when\ g^{[l]}(z) = ReLU(z) \\ np.random.randn(shape) * np.sqrt(\frac{1}{n^{[l-1]}}),\ when\ g^{[l]}(z) = tanh(z) \end{cases}$