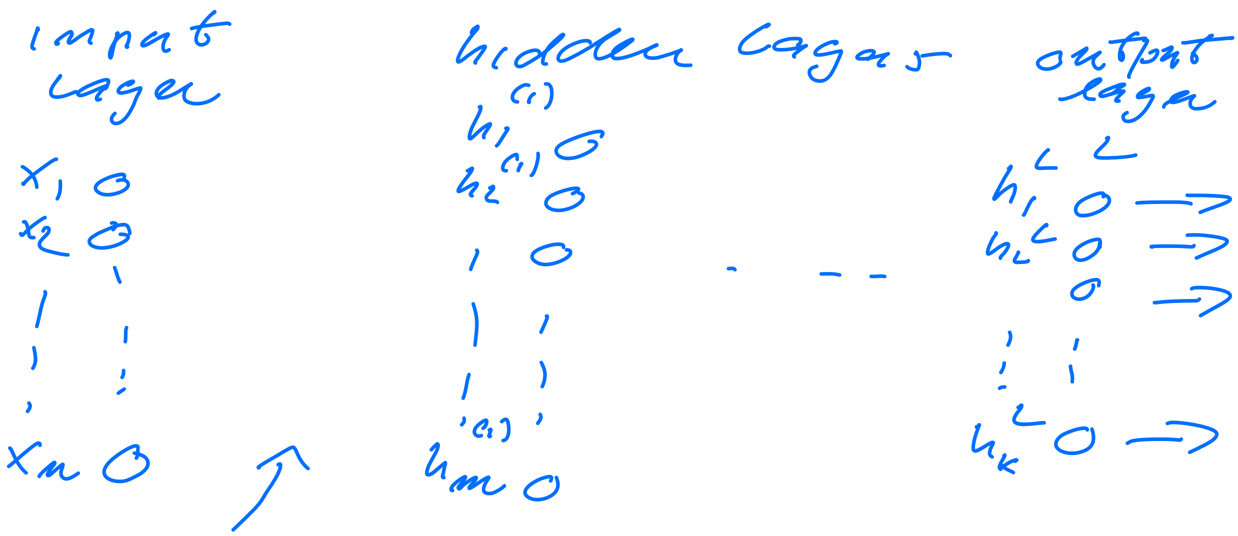
Lecture December 13

The mathematics of NN



$$W \in \mathbb{R}^{n \times m}$$

input to - h -

$$z(x) = W^T x + b$$

$$\sigma(z) = \underbrace{[a_1(x), a_2(x) \dots a_m(x)]}_{A}$$

with many hidden layers

$$1 \le l \le L$$

an Artificial NN cascades the operations $\sigma(z)$ multiple times

$$\sigma_L(A_L(\cdots \sigma_1(A_1(x))))$$

consider a simple NN in which w and b are scalars $L=2$

$$f(x;\theta) = \sigma_2(w_2\sigma_1(w_1x+b_1) + b_2)$$

$$a_1 = \sigma_1(w_1x+b_1)$$

input to L is $a_1 w_2 + b_2$

partial derivatives wrt w_1

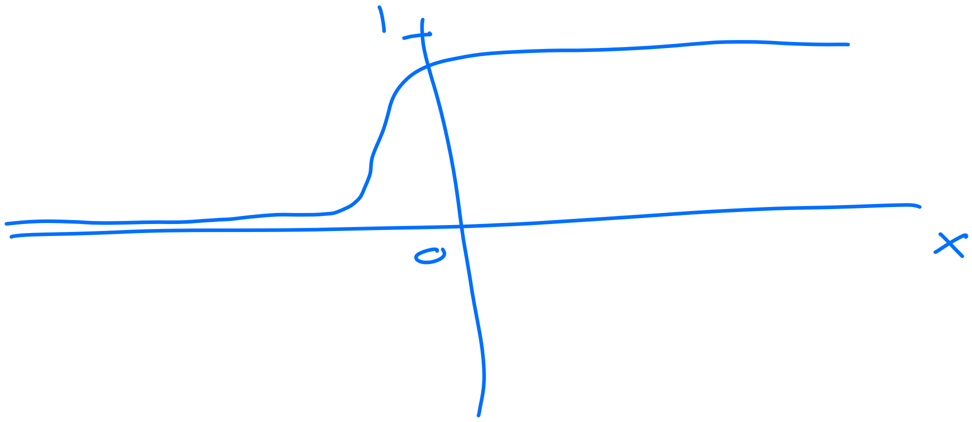
$$\partial_{w_1} f(x;\theta) = \sigma_2'(w_2\sigma_1(w_1x+b_1)+b_2) \times w_2\sigma_1'(w_1x+b_1)x$$

with L-layers

$$\partial_{w_1} f(x;\theta) = \left[\prod_{l=2}^{L} w_l\right] \times \left[\prod_{l=1}^{L} \sigma_l'(z_l)\right] x$$

$$z_\ell = A_\ell(\sigma_{\ell-1}(A_{\ell-1}(\cdots \sigma_1(A_1(x)))))$$

if σ_ℓ is the sigmoid function (or tanh)



$\sigma_\ell'(x)$ will be small if $|x| \gg 0$ $\Rightarrow$ vanishing gradients

Typical activation functions

- ReLU function

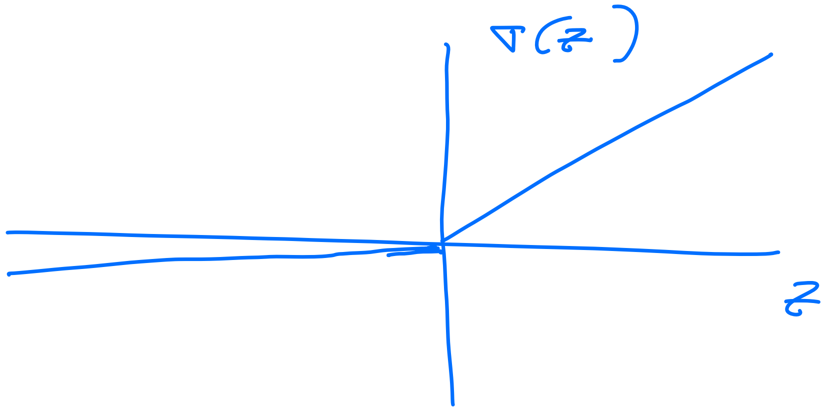
$$\sigma(z) = \max(0, z)$$

$$\sigma'(z) = 1 \quad \text{for } z > 0$$

- Leaky ReLU

$$\sigma(z) = \begin{cases} \alpha \cdot z & z < 0 \end{cases}$$

$$\sigma(z) = \begin{cases} \ldots & \\ z & \geq 0 \end{cases}$$



$$\alpha = 0.01$$

$$\sigma'(z) = \begin{cases} \alpha & z < 0 \\ 1 & z \geq 0 \end{cases}$$

$$ELU = \sigma(z) = \begin{cases} \alpha(e^{z} - 1) & z \leq 0 \\ z & \text{for } > 0 \end{cases}$$

$$\sigma'(z) = \begin{cases} \alpha e^{z} & z \leq 0 \\ 1 \text{ for } & z > 0 \end{cases}$$

$$\sigma(z) = \tanh z \qquad z \in (-1, 1)$$

$$\sigma'(z) = 1 - (\tanh z)^2$$

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

$$\sigma'(z) = \sigma(z)\big(1 - \sigma(z)\big)$$