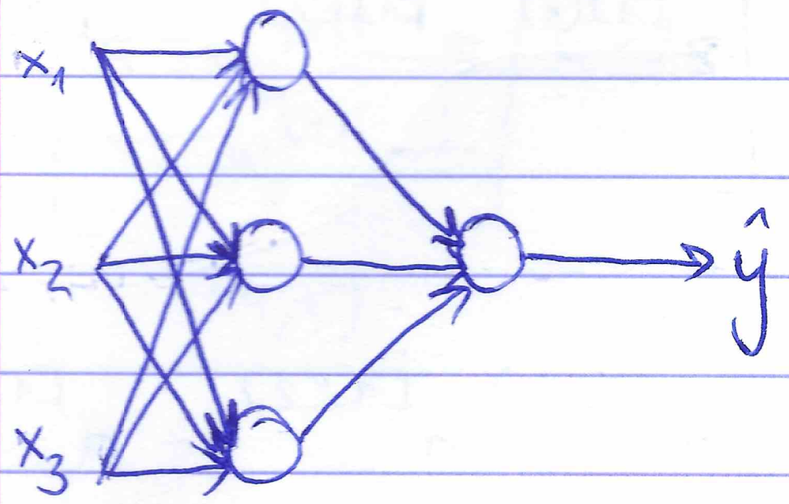
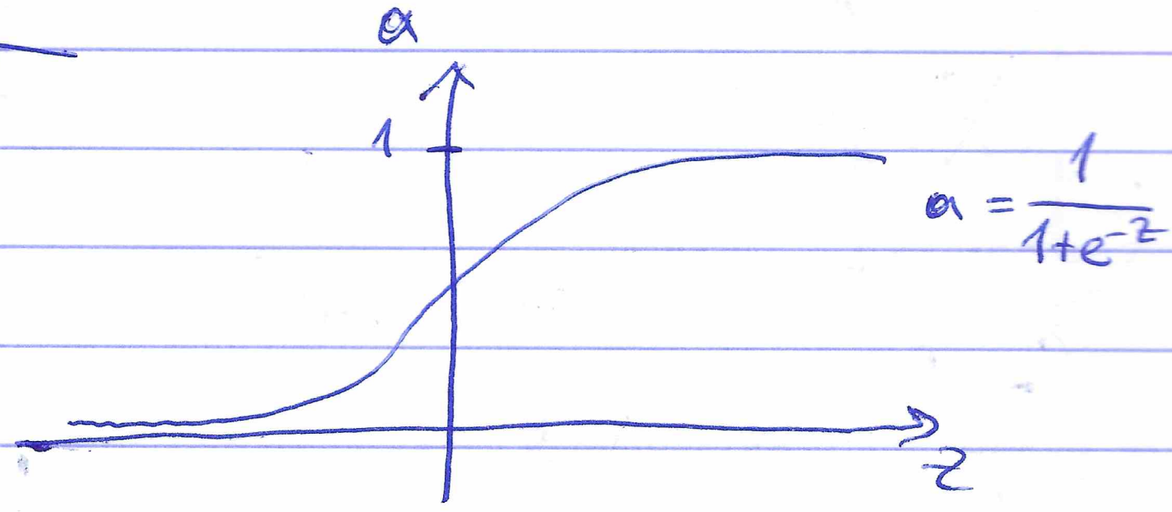
Activation Functions



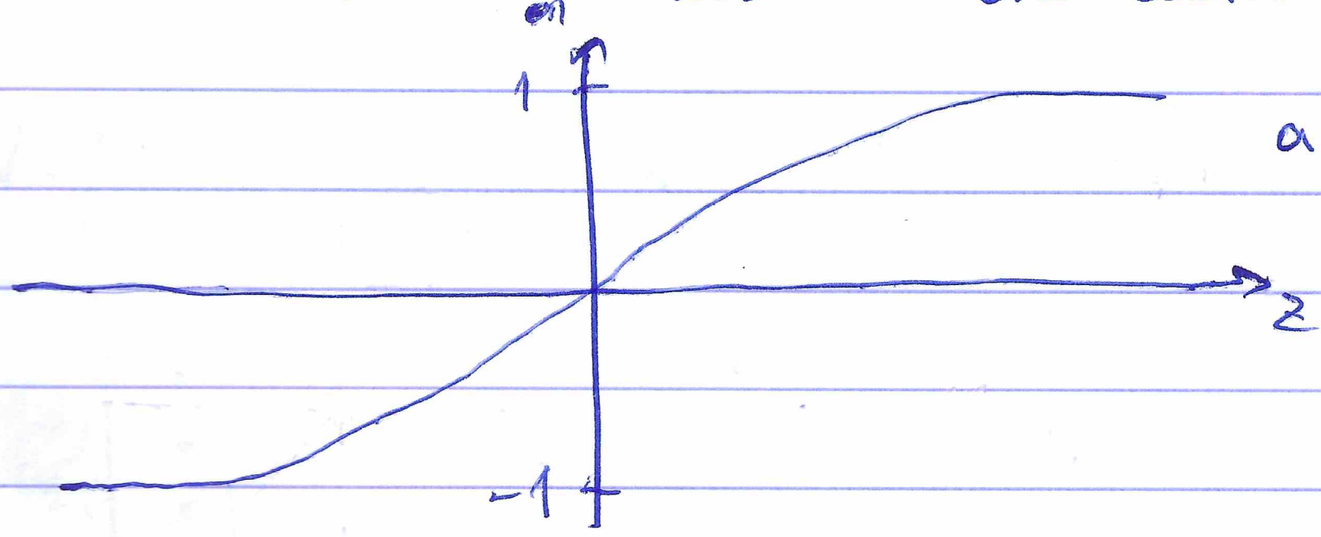


Given x:

$$z^{[1]} = W^{[1]} x + b^{[1]}$$
$$a^{[1]} = \sigma(z^{[1]}) \rightarrow g(z^{[1]})$$
$$z^{[2]} = W^{[2]} a^{[1]} + b^{[2]}$$
$$a^{[2]} = \sigma(z^{[2]}) \rightarrow g(z^{[2]})$$

In the more general case, we can have $g(z)$, which could be a non-linear function that may not be the sigmoid function

An activation function that works better is the tanh function:



$a = \tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$

(a shifted sigmoid)

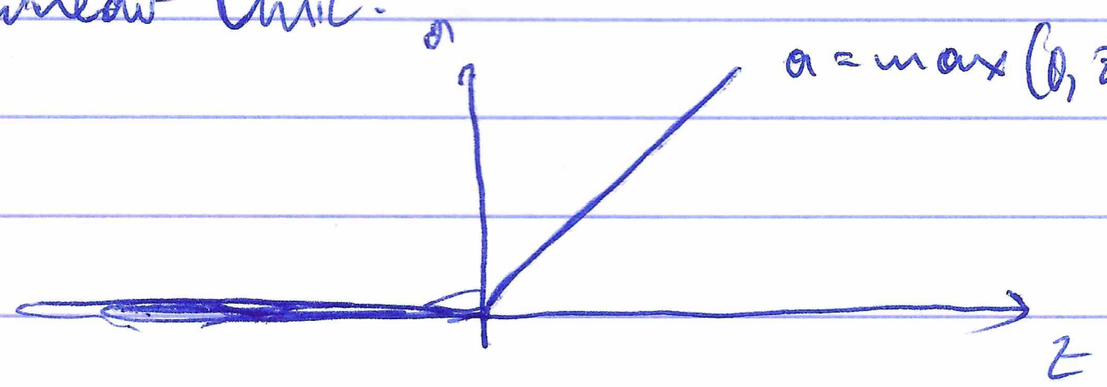
So, $g(z^{[1]}) = \tanh(z^{[1]})$ works better. It also has the effect of centering the data (mean ≈ 0), which makes learning for the next layer a bit easier.

The exception is for the output layer. If $y \in \{0, 1\}$, then $0 \leq \hat{y} \leq 1$

$g^{[2]}(z^{[2]}) = \sigma(z^{[2]})$

-1 1

Rectified Linear Unit:



Leaky ReLU