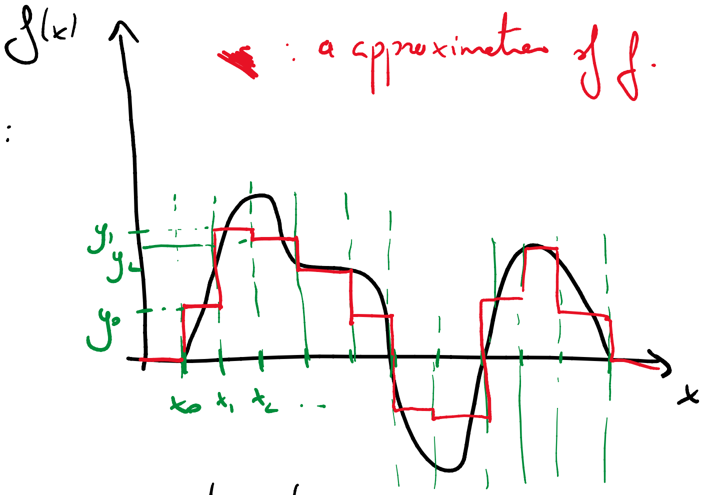
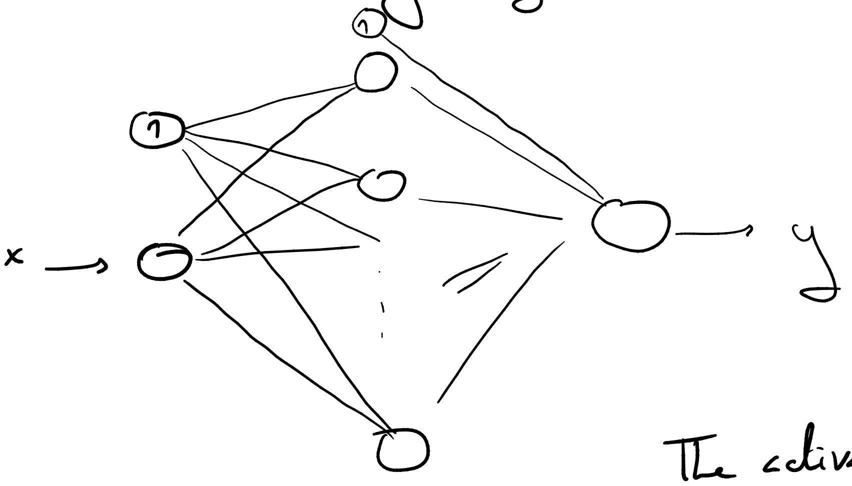
A: Neural-network with one hidden layer are **universal approximator**

Let's see a simple example:



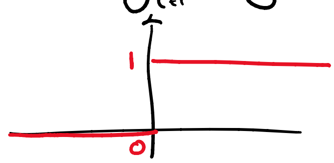
Let's consider the following neural-net.



- 1 input x
- Nh hidden nodes on 1st hidden layer
- 1 output y

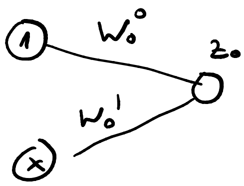
The activation fct:

- For the hidden layer: $\Theta(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{if } x < 0 \end{cases}$ (heaviside)



- For the output: linear activation $f(x) = x$

Let focus on the first hidden node



I want to have: $z_1 = 1$ if $x > x_0$

0 otherwise

I want to have: $z_1 = 1$ if $x > x_0$

0 otherwise

$w_0^0 + x_0 w_0^1 = 0$

$$\rightarrow \boxed{\frac{w_0^1}{w_0^0} = x_0}$$

we want the output to be y_0 when z_0 is activated!

$$y = w_y^0 + w_y^1 \cdot z_1 = y_0 \qquad (z_1 = 1)$$

$$\rightarrow \begin{cases} w_y^0 = 0 \\ w_y^1 = y_0 \end{cases}$$

Let's look at the second hidden node.

activated for $x > x_1$

$$\rightarrow \boxed{\frac{w_1^1}{w_1^0} = x_1}$$

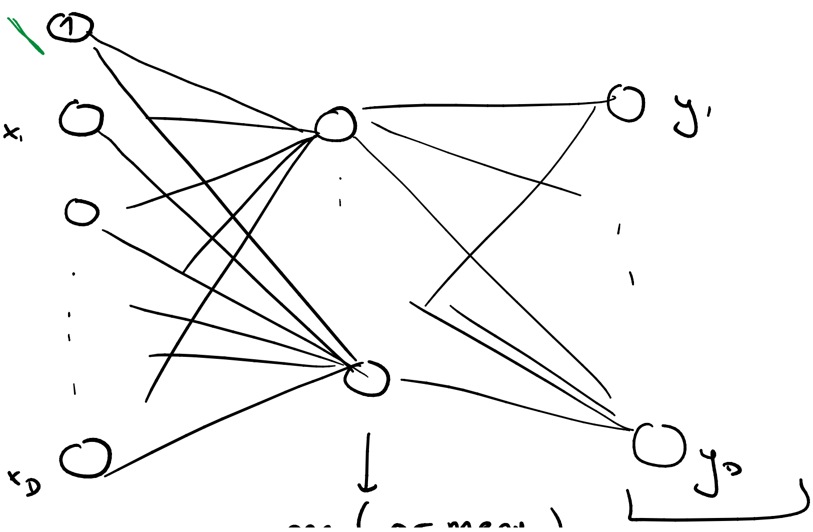
for $x_1 < x < x_2$

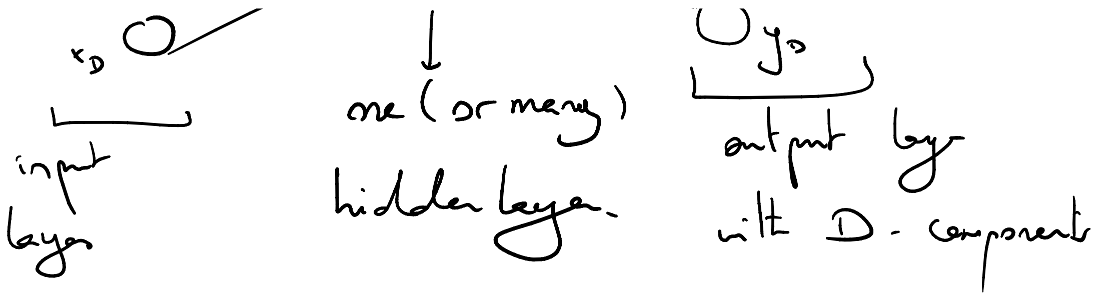
$$y = w_y^0 + w_y^1 + w_y^2 = y_1$$

($w_y^0 \to 0$, $w_y^1 \to y_0$)

$$\boxed{w_y^2 = y_1 - y_0}$$

B) Autoencoders (AE)





It is unsupervised : you do not need a label.
(self-supervised)

In its simplest form (all linear activation functions)

the solution of the square-loss is equivalent to a decomposition in principal components.

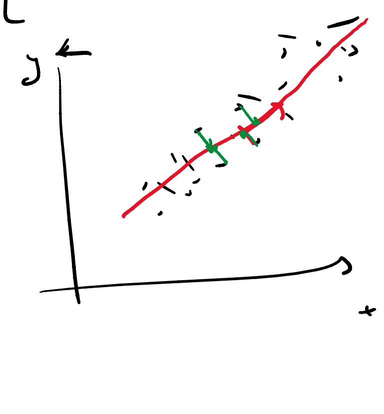
- PCA on a small number k of sub-spaces

→ it finds the subspace in $\mathbb{R}^n$ such that

[the error of the on the projection of the dataset on this subspace is the smallest possible one according to the L_2-norm

⇕

[the variance of the dataset projected into this subspace is the highest possible one



The loss of this simple AE: $\mathcal{L} = \| \vec{x} - \vec{y} \|$

$$\vec{y} = W^{(1)'} W^{(1)} \vec{x} \qquad \mathcal{L} = \| \vec{x} - W^{(1)'} W^{(1)} . \vec{x} \|^2$$

a solution: $\begin{cases} W^{(1)'} = W^{(1)T} \\ W^{(1)} : \text{set of } M_h \text{ principal vectors} \end{cases}$

⚠ activation fct on the output!

if $\begin{cases} \vec{x} \in \mathbb{R}^D \to \text{you use the linear activation and } L_2\text{-norm} \\ \vec{x} \in [0,1] \to \text{it might better to use a sigmoid act fct.} \end{cases}$

with the cross-entropy loss

$$\mathcal{L} = \sum_i \left[x_i \log y_i + (1-x_i) \log(1-y_i) \right]$$

$$\left(y_i^{x_i} (1-y_i)^{1-x_i} \right)$$

Lab Work: . how to encode digits.

. how we can create a denoiser

. ______________ a machine that to a given digit associate the next one.