# EE-559 – Deep learning

# 3.4. Multi-Layer Perceptrons

François Fleuret

https://fleuret.org/ee559/

Wed Nov 21 08:27:39 UTC 2018

RESEARCH INSTITUTE

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

So far we have seen linear classifiers of the form

$$\begin{aligned} \mathbb{R}^D &\to \mathbb{R} \\ x &\mapsto \sigma(w \cdot x + b), \end{aligned}$$

with $w \in \mathbb{R}^D$, $b \in \mathbb{R}$, and $\sigma : \mathbb{R} \to \mathbb{R}$.

So far we have seen linear classifiers of the form

$$\begin{aligned}\mathbb{R}^D &\to \mathbb{R} \\ x &\mapsto \sigma(w \cdot x + b),\end{aligned}$$

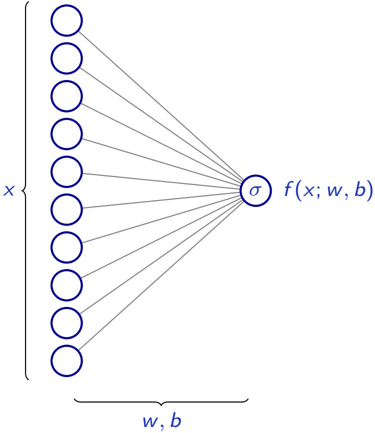
with $w \in \mathbb{R}^D$, $b \in \mathbb{R}$, and $\sigma : \mathbb{R} \to \mathbb{R}$.

This can naturally be extended to a multi-dimension output by applying a similar transformation to every output, which leads to

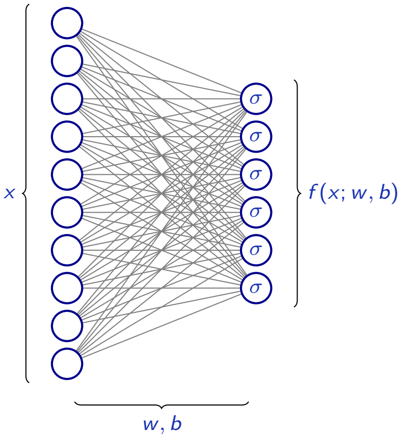
$$\begin{aligned}\mathbb{R}^D &\to \mathbb{R}^C \\ x &\mapsto \sigma(wx + b),\end{aligned}$$

with $w \in \mathbb{R}^{C \times D}$, $b \in \mathbb{R}^C$, and $\sigma$ is applied component-wise.

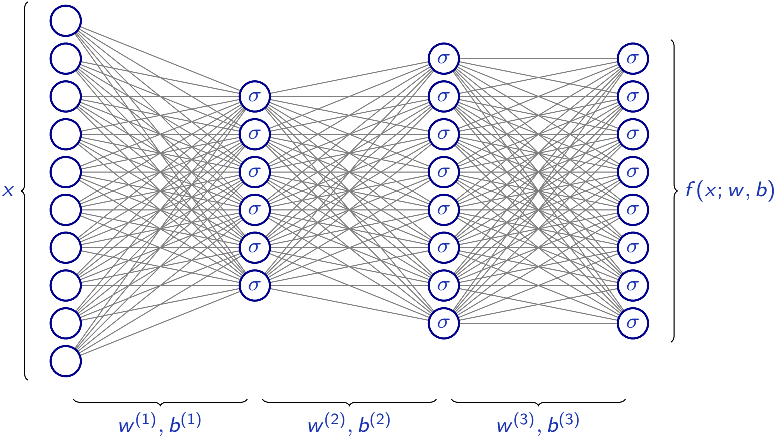
Even though it has no practical value implementation-wise, we can represent such a model as a combination of units, and extend it.



Even though it has no practical value implementation-wise, we can represent such a model as a combination of units, and extend it.



Even though it has no practical value implementation-wise, we can represent such a model as a combination of units, and extend it.



This latter structure can be formally defined, with $x^{(0)} = x$,

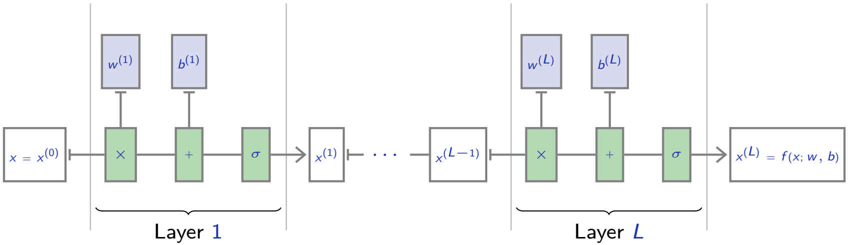
$$\forall l = 1, \dots, L,\ x^{(l)} = \sigma\left(w^{(l)} x^{(l-1)} + b^{(l)}\right)$$

and $f(x; w, b) = x^{(L)}$.

This latter structure can be formally defined, with $x^{(0)} = x$,

$$\forall l = 1, \dots, L, \; x^{(l)} = \sigma\left(w^{(l)} x^{(l-1)} + b^{(l)}\right)$$

and $f(x; w, b) = x^{(L)}$.




Such a model is a **Multi-Layer Perceptron (MLP).**

Note that if $\sigma$ is an affine transformation, the full MLP is a composition of affine mappings, and itself an affine mapping.

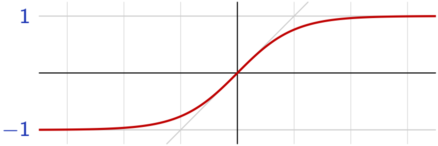
Note that if $\sigma$ is an affine transformation, the full MLP is a composition of affine mappings, and itself an affine mapping.

Consequently:

⚠ **The activation function $\sigma$ should be non-linear**, or the resulting MLP is an affine mapping with a peculiar parametrization.

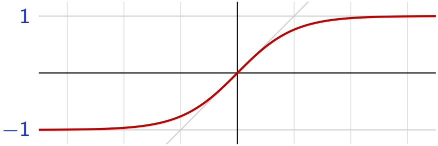
The two classical activation functions are the hyperbolic tangent

$$x \mapsto \frac{2}{1+e^{-2x}} - 1$$



The two classical activation functions are the hyperbolic tangent

$$x \mapsto \frac{2}{1 + e^{-2x}} - 1$$

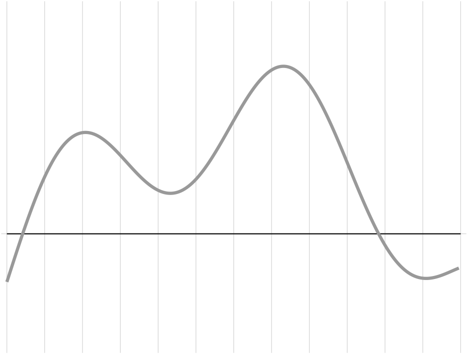



and the rectified linear unit (ReLU)

$$x \mapsto \max(0, x)$$

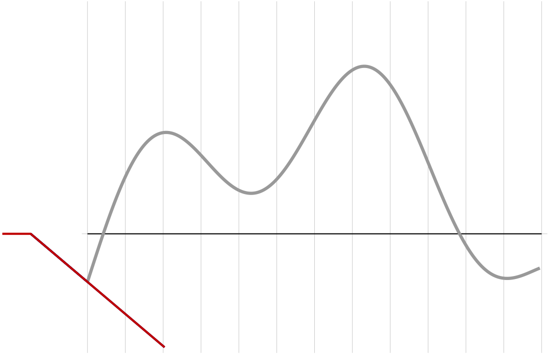
# Universal approximation

We can approximate any $\psi \in \mathscr{C}([a, b], \mathbb{R})$ with a linear combination of translated/scaled ReLU functions.



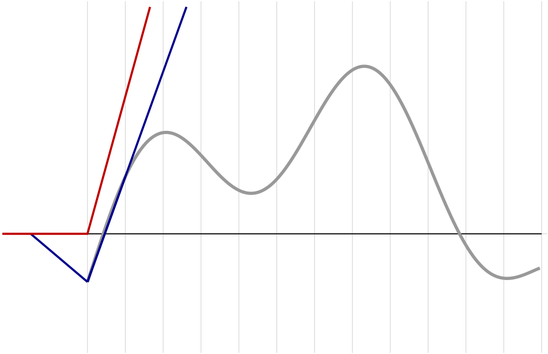
We can approximate any $\psi \in \mathscr{C}([a, b], \mathbb{R})$ with a linear combination of translated/scaled ReLU functions.

$$f(x) = \sigma(w_1 x + b_1)$$



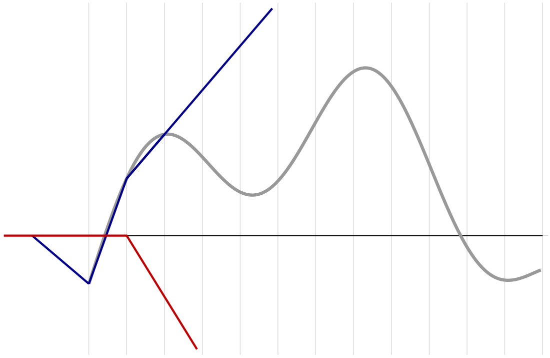
We can approximate any $\psi \in \mathscr{C}([a, b], \mathbb{R})$ with a linear combination of translated/scaled ReLU functions.

$$f(x) = \sigma(w_1 x + b_1) + \sigma(w_2 x + b_2)$$



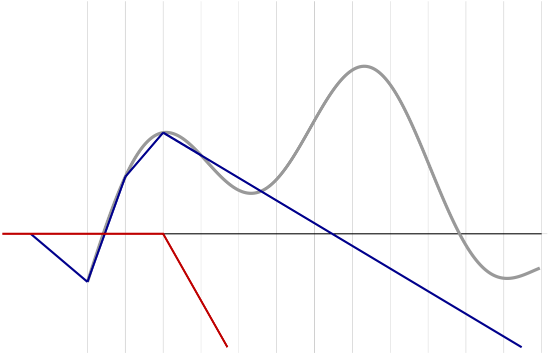
We can approximate any $\psi \in \mathscr{C}([a, b], \mathbb{R})$ with a linear combination of translated/scaled ReLU functions.

$$f(x) = \sigma(w_1 x + b_1) + \sigma(w_2 x + b_2) + \sigma(w_3 x + b_3)$$



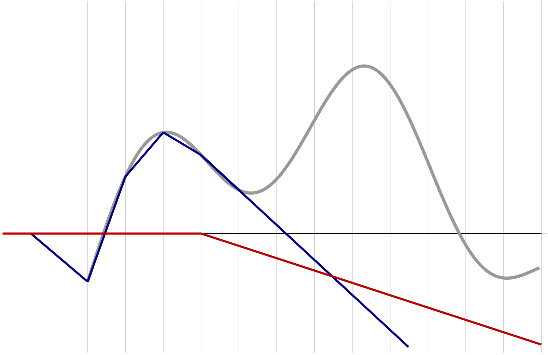
We can approximate any $\psi \in \mathscr{C}([a, b], \mathbb{R})$ with a linear combination of translated/scaled ReLU functions.

$$f(x) = \sigma(w_1 x + b_1) + \sigma(w_2 x + b_2) + \sigma(w_3 x + b_3) + \dots$$



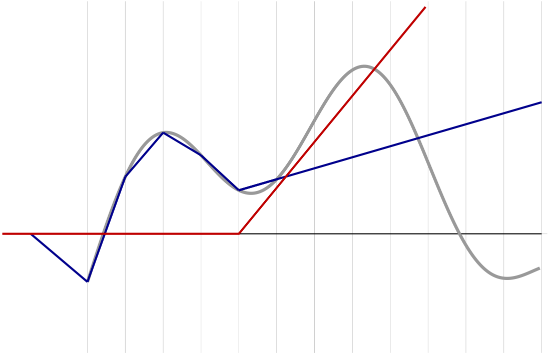
We can approximate any $\psi \in \mathscr{C}([a, b], \mathbb{R})$ with a linear combination of translated/scaled ReLU functions.

$$f(x) = \sigma(w_1 x + b_1) + \sigma(w_2 x + b_2) + \sigma(w_3 x + b_3) + \dots$$



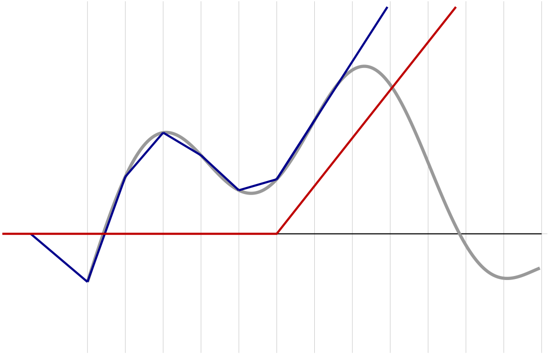
We can approximate any $\psi \in \mathscr{C}([a, b], \mathbb{R})$ with a linear combination of translated/scaled ReLU functions.

$$f(x) = \sigma(w_1 x + b_1) + \sigma(w_2 x + b_2) + \sigma(w_3 x + b_3) + \dots$$



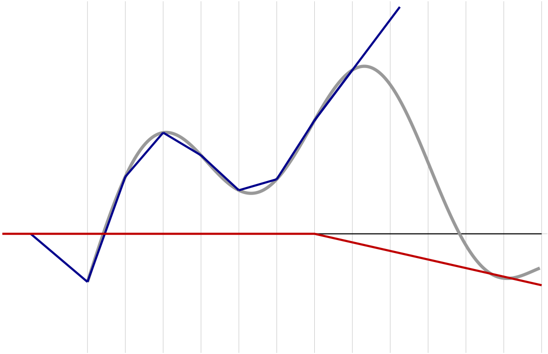
We can approximate any $\psi \in \mathscr{C}([a, b], \mathbb{R})$ with a linear combination of translated/scaled ReLU functions.

$$f(x) = \sigma(w_1 x + b_1) + \sigma(w_2 x + b_2) + \sigma(w_3 x + b_3) + \dots$$



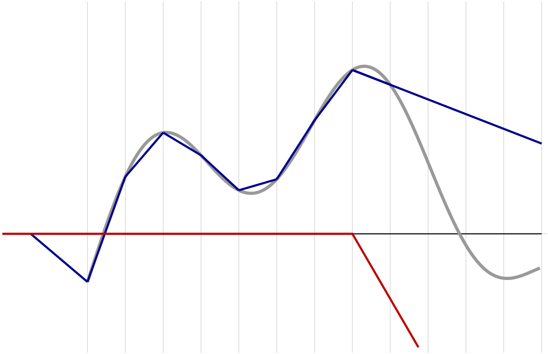
We can approximate any $\psi \in \mathscr{C}([a, b], \mathbb{R})$ with a linear combination of translated/scaled ReLU functions.

$$f(x) = \sigma(w_1 x + b_1) + \sigma(w_2 x + b_2) + \sigma(w_3 x + b_3) + \dots$$



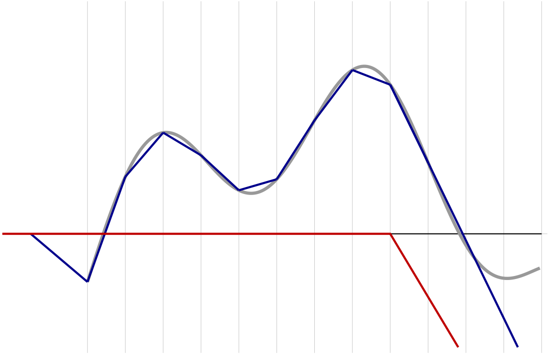
We can approximate any $\psi \in \mathscr{C}([a, b], \mathbb{R})$ with a linear combination of translated/scaled ReLU functions.

$$f(x) = \sigma(w_1 x + b_1) + \sigma(w_2 x + b_2) + \sigma(w_3 x + b_3) + \dots$$



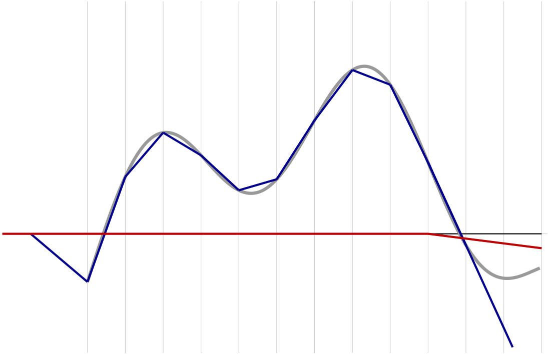
We can approximate any $\psi \in \mathscr{C}([a, b], \mathbb{R})$ with a linear combination of translated/scaled ReLU functions.

$$f(x) = \sigma(w_1 x + b_1) + \sigma(w_2 x + b_2) + \sigma(w_3 x + b_3) + \dots$$



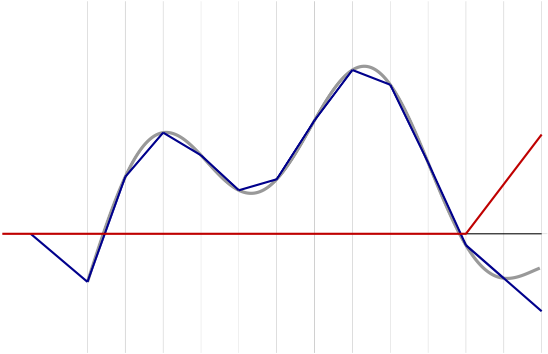
We can approximate any $\psi \in \mathscr{C}([a, b], \mathbb{R})$ with a linear combination of translated/scaled ReLU functions.

$$f(x) = \sigma(w_1 x + b_1) + \sigma(w_2 x + b_2) + \sigma(w_3 x + b_3) + \dots$$



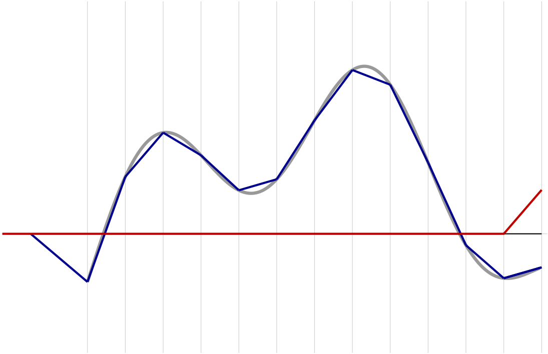
We can approximate any $\psi \in \mathscr{C}([a, b], \mathbb{R})$ with a linear combination of translated/scaled ReLU functions.

$$f(x) = \sigma(w_1 x + b_1) + \sigma(w_2 x + b_2) + \sigma(w_3 x + b_3) + \dots$$



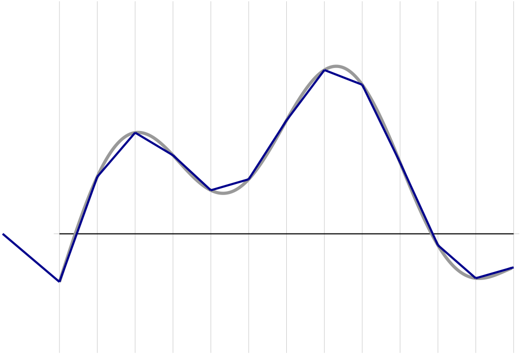
We can approximate any $\psi \in \mathscr{C}([a, b], \mathbb{R})$ with a linear combination of translated/scaled ReLU functions.

$$f(x) = \sigma(w_1 x + b_1) + \sigma(w_2 x + b_2) + \sigma(w_3 x + b_3) + \dots$$



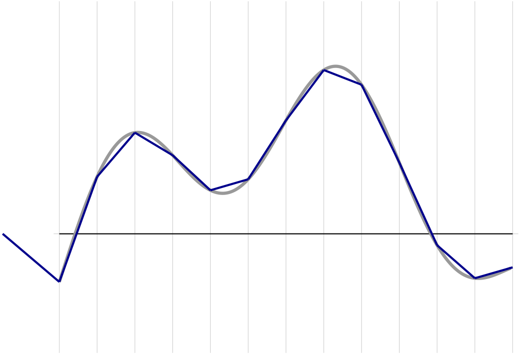
We can approximate any $\psi \in \mathscr{C}([a, b], \mathbb{R})$ with a linear combination of translated/scaled ReLU functions.

$$f(x) = \sigma(w_1 x + b_1) + \sigma(w_2 x + b_2) + \sigma(w_3 x + b_3) + \dots$$



We can approximate any $\psi \in \mathscr{C}([a, b], \mathbb{R})$ with a linear combination of translated/scaled ReLU functions.

$$f(x) = \sigma(w_1 x + b_1) + \sigma(w_2 x + b_2) + \sigma(w_3 x + b_3) + \dots$$



This is true for other activation functions under mild assumptions.

Extending this result to any $\psi \in \mathscr{C}([0,1]^D, \mathbb{R})$ requires a bit of work.

Extending this result to any $\psi \in \mathscr{C}([0,1]^D, \mathbb{R})$ requires a bit of work.

First, we can use the previous result for the sin function

$$\forall A > 0, \epsilon > 0, \ \exists N, \ (\alpha_n, a_n) \in \mathbb{R} \times \mathbb{R}, n = 1, \dots, N,$$

$$\text{s.t.} \ \max_{x \in [-A, A]} \left| \sin(x) - \sum_{n=1}^{N} \alpha_n \sigma(x - a_n) \right| \leq \epsilon.$$

Extending this result to any $\psi \in \mathscr{C}([0,1]^D, \mathbb{R})$ requires a bit of work.

First, we can use the previous result for the sin function

$$\forall A > 0, \epsilon > 0,\ \exists N,\ (\alpha_n, a_n) \in \mathbb{R} \times \mathbb{R}, n = 1, \dots, N,$$
$$\text{s.t.} \max_{x \in [-A,A]} \left| \sin(x) - \sum_{n=1}^{N} \alpha_n \sigma(x - a_n) \right| \leq \epsilon.$$

And the density of Fourier series provides

$$\forall \psi \in \mathscr{C}([0,1]^D, \mathbb{R}), \delta > 0, \exists M, (v_m, \gamma_m, c_m) \in \mathbb{R}^D \times \mathbb{R} \times \mathbb{R}, m = 1, \dots, M,$$
$$\text{s.t.} \max_{x \in [0,1]^D} \left| \psi(x) - \sum_{m=1}^{M} \gamma_m \sin(v_m \cdot x + c_m) \right| \leq \delta.$$

So, $\forall \xi > 0$, with

$$\delta = \frac{\xi}{2}, A = \max_{1 \leq m \leq M} \max_{x \in [0,1]^D} \left| v_m \cdot x + c_m \right|, \text{ and } \epsilon = \frac{\xi}{2 \sum_m |\gamma_m|}$$

we get, $\forall x \in [0,1]^D$,

$$\left| \psi(x) - \sum_{m=1}^{M} \gamma_m \left( \sum_{n=1}^{N} \alpha_n \sigma(v_m \cdot x + c_m - a_n) \right) \right|$$

$$\leq \underbrace{\left| \psi(x) - \sum_{m=1}^{M} \gamma_m \sin(v_m \cdot x + c_m) \right|}_{\leq \frac{\xi}{2}}$$

$$\underbrace{+ \sum_{m=1}^{M} |\gamma_m| \underbrace{\left| \sin(v_m \cdot x + c_m) - \sum_{n=1}^{N} \alpha_n \sigma(v_m \cdot x + c_m - a_n) \right|}_{\leq \frac{\xi}{2 \sum_m |\gamma_m|}}}_{\leq \frac{\xi}{2}}$$

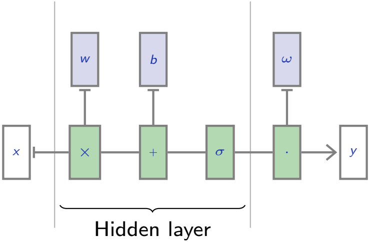
So we can approximate any continuous function

$$\psi : [0,1]^D \to \mathbb{R}$$

with a one hidden layer perceptron

$$x \mapsto \omega \cdot \sigma(w\,x + b),$$

where $b \in \mathbb{R}^K$, $w \in \mathbb{R}^{K \times D}$, and $\omega \in \mathbb{R}^K$.



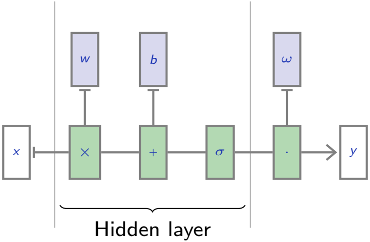
So we can approximate any continuous function

$$\psi : [0,1]^D \to \mathbb{R}$$

with a one hidden layer perceptron

$$x \mapsto \omega \cdot \sigma(w\,x + b),$$

where $b \in \mathbb{R}^K$, $w \in \mathbb{R}^{K \times D}$, and $\omega \in \mathbb{R}^K$.




This is the **universal approximation theorem.**

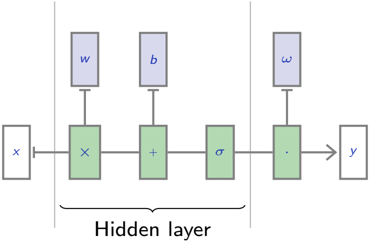
So we can approximate any continuous function

$$\psi : [0,1]^D \to \mathbb{R}$$

with a one hidden layer perceptron

$$x \mapsto \omega \cdot \sigma(w\,x + b),$$

where $b \in \mathbb{R}^K$, $w \in \mathbb{R}^{K \times D}$, and $\omega \in \mathbb{R}^K$.




This is the **universal approximation theorem.**

⚠ A better approximation requires a larger hidden layer (larger $K$), and this theorem says nothing about the relation between the two.

The end