EE-559 – Deep learning

3.4. Multi-Layer Perceptrons

François Fleuret

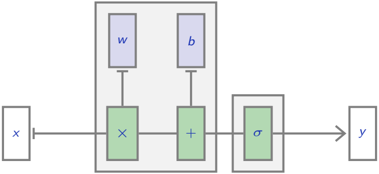
https://fleuret.org/ee559/

Tue Oct 30 07:08:22 UTC 2018

RESEARCH INSTITUTE

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

For flexibility, we will separate the linear operators and the non-linearities in different blocks in our figures.



We can combine several "layers":

With $x^{(0)} = x$,

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

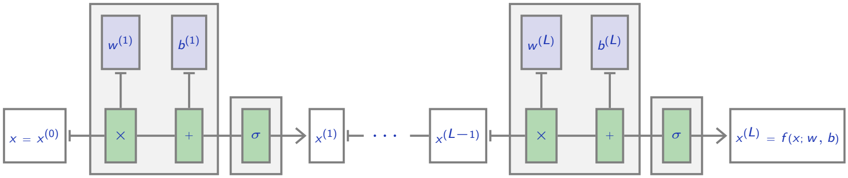
and $f(x; w, b) = x^{(L)}$.

We can combine several "layers":

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 σ is a linear transformation,

$$\forall x \in \mathbb{R}^N,\ \sigma(x) = \alpha x + \beta \mathbb{I}$$

with $\alpha, \beta \in \mathbb{R}$, we have

$$\forall l = 1, \dots, L,\ x^{(l)} = \alpha w^{(l)} x^{(l-1)} + \alpha b^{(l)} + \beta \mathbb{I},$$

Note that if σ is a linear transformation,

$$\forall x \in \mathbb{R}^N,\ \sigma(x) = \alpha x + \beta \mathbb{I}$$

with $\alpha, \beta \in \mathbb{R}$, we have

$$\forall l = 1, \dots, L,\ x^{(l)} = \alpha w^{(l)} x^{(l-1)} + \alpha b^{(l)} + \beta \mathbb{I},$$

and the whole mapping is an affine transform

$$f(x; w, b) = A^{(L)} x + B^{(L)}$$

where $A^{(0)} = \mathbb{I}, B^{(0)} = 0$ and

$$\forall l < L,\ \begin{cases} A^{(l)} = \alpha\, w^{(l)} A^{(l-1)} \\ B^{(l)} = \alpha\, w^{(l)} B^{(l-1)} + \alpha\, b^{(l)} + \beta \mathbb{I} \end{cases}$$

Note that if σ is a linear transformation,

$$\forall x \in \mathbb{R}^N,\ \sigma(x) = \alpha x + \beta \mathbb{I}$$

with $\alpha, \beta \in \mathbb{R}$, we have

$$\forall l = 1, \dots, L,\ x^{(l)} = \alpha w^{(l)} x^{(l-1)} + \alpha b^{(l)} + \beta \mathbb{I},$$

and the whole mapping is an affine transform

$$f(x; w, b) = A^{(L)} x + B^{(L)}$$

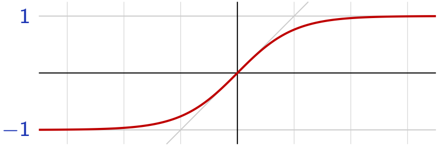
where $A^{(0)} = \mathbb{I}$, $B^{(0)} = 0$ and

$$\forall l < L,\ \begin{cases} A^{(l)} = \alpha\, w^{(l)} A^{(l-1)} \\ B^{(l)} = \alpha\, w^{(l)} B^{(l-1)} + \alpha\, b^{(l)} + \beta \mathbb{I} \end{cases}$$

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

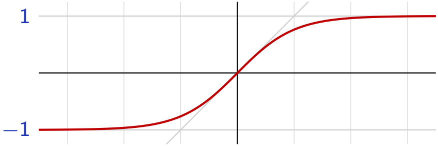
The two classical activation functions are the hyperbolic tangent

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



The two classical activation functions are the hyperbolic tangent

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

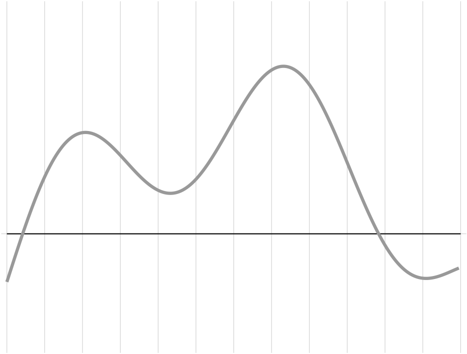


and the rectified linear unit (ReLU)

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

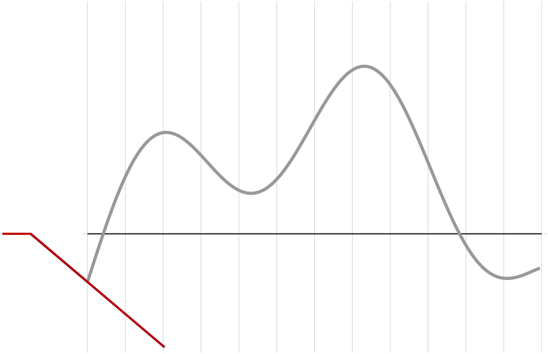
Universal approximation

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



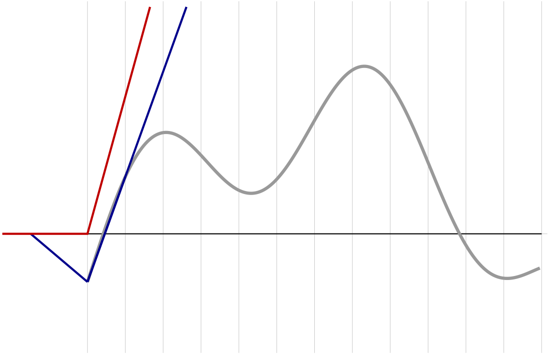
We can approximate any $\psi \in \mathscr{C}([a, b], \mathbb{R})$ with a linear combination of translated/scaled ReLU functions.

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



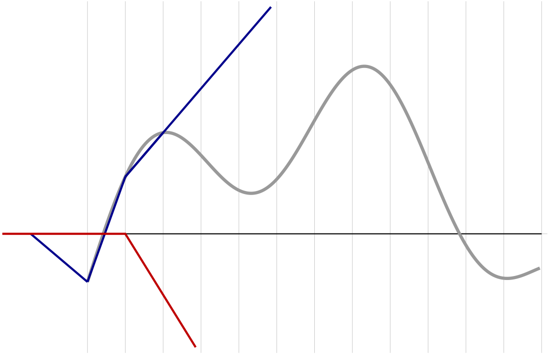
We can approximate any $\psi \in \mathscr{C}([a, b], \mathbb{R})$ with a linear combination of translated/scaled ReLU functions.

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



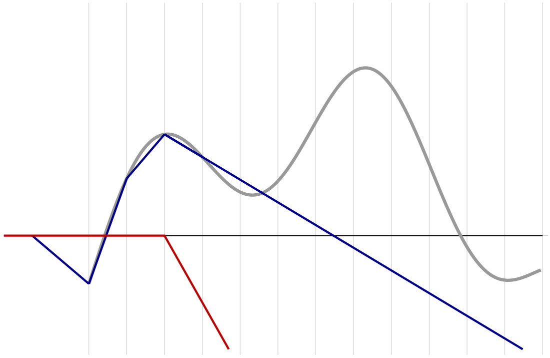
We can approximate any $\psi \in \mathscr{C}([a, b], \mathbb{R})$ with a linear combination of translated/scaled ReLU functions.

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



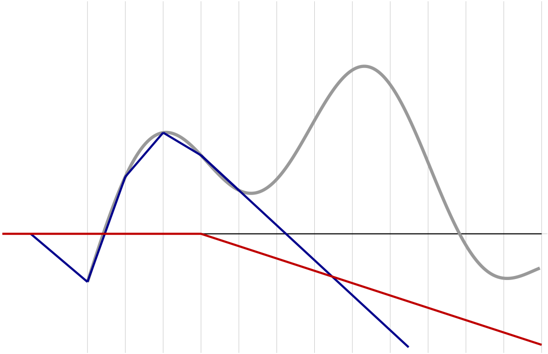
We can approximate any $\psi \in \mathscr{C}([a, b], \mathbb{R})$ with a linear combination of translated/scaled ReLU functions.

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



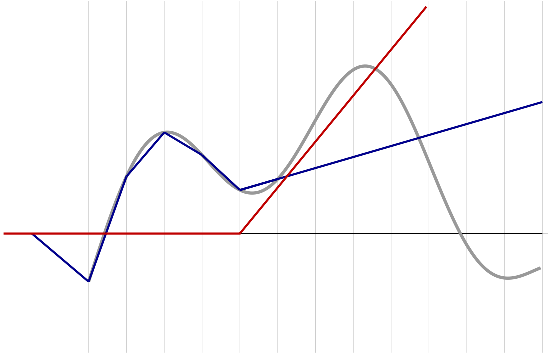
We can approximate any $\psi \in \mathscr{C}([a, b], \mathbb{R})$ with a linear combination of translated/scaled ReLU functions.

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



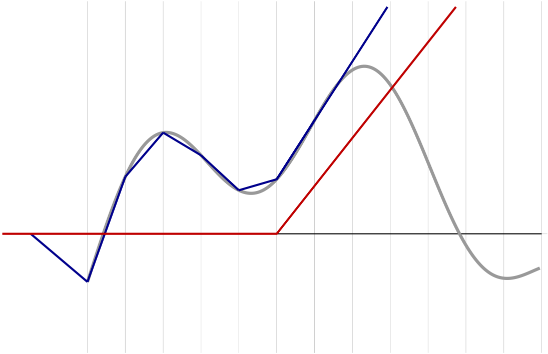
We can approximate any $\psi \in \mathscr{C}([a, b], \mathbb{R})$ with a linear combination of translated/scaled ReLU functions.

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



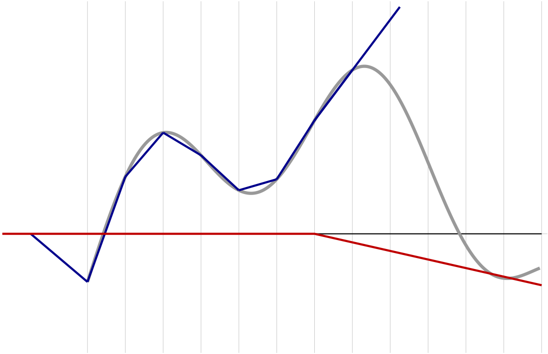
We can approximate any $\psi \in \mathscr{C}([a, b], \mathbb{R})$ with a linear combination of translated/scaled ReLU functions.

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



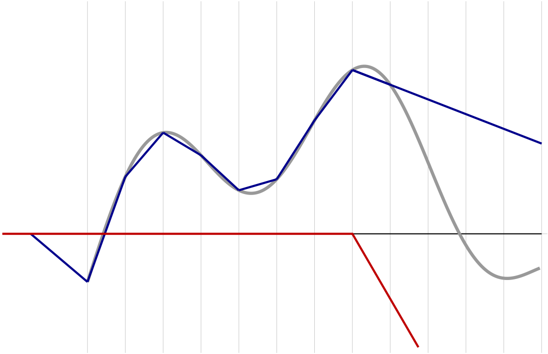
We can approximate any $\psi \in \mathscr{C}([a, b], \mathbb{R})$ with a linear combination of translated/scaled ReLU functions.

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



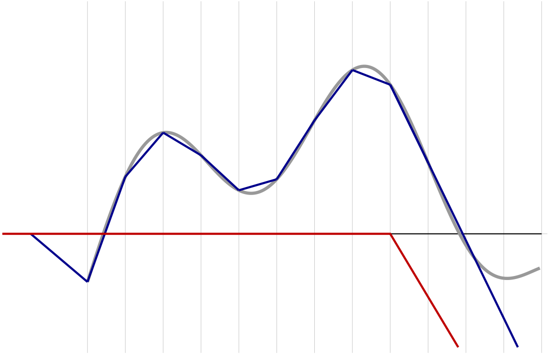
We can approximate any $\psi \in \mathscr{C}([a, b], \mathbb{R})$ with a linear combination of translated/scaled ReLU functions.

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



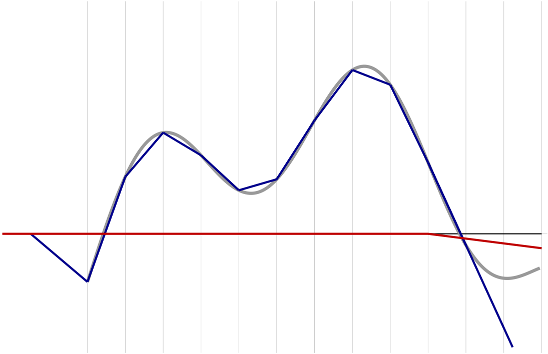
We can approximate any $\psi \in \mathscr{C}([a, b], \mathbb{R})$ with a linear combination of translated/scaled ReLU functions.

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



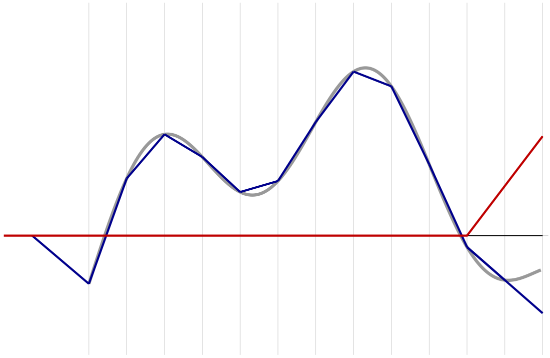
We can approximate any $\psi \in \mathscr{C}([a,b],\mathbb{R})$ with a linear combination of translated/scaled ReLU functions.

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



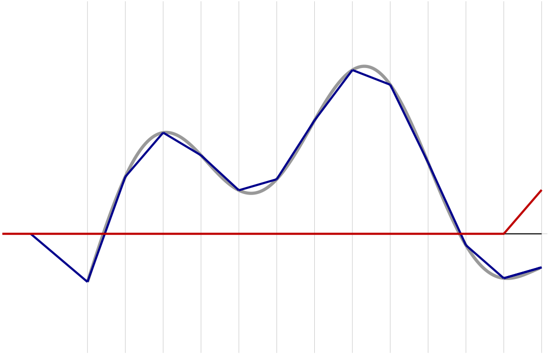
We can approximate any $\psi \in \mathscr{C}([a, b], \mathbb{R})$ with a linear combination of translated/scaled ReLU functions.

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



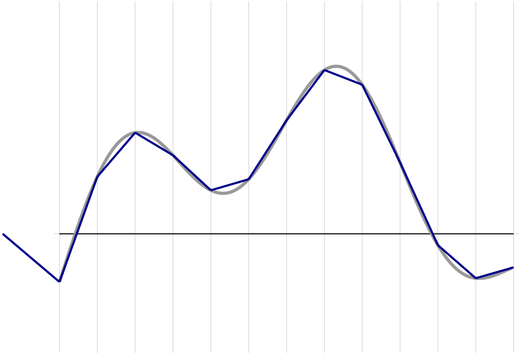
We can approximate any $\psi \in \mathscr{C}([a, b], \mathbb{R})$ with a linear combination of translated/scaled ReLU functions.

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



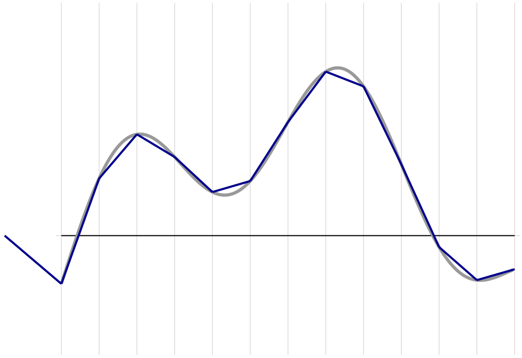
We can approximate any $\psi \in \mathscr{C}([a, b], \mathbb{R})$ with a linear combination of translated/scaled ReLU functions.

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



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

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

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

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

So we can approximate any continuous function

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

with a mapping of the form

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

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

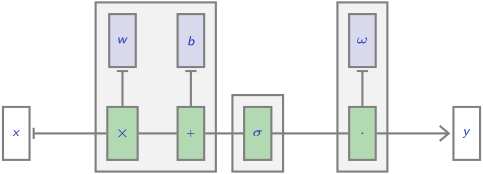
So we can approximate any continuous function

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

with a mapping of the form

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

where $b \in \mathbb{R}^K$, $w \in \mathbb{R}^{K \times D}$, and $\omega \in \mathbb{R}^K$, *i.e.* with a one hidden layer perceptron.



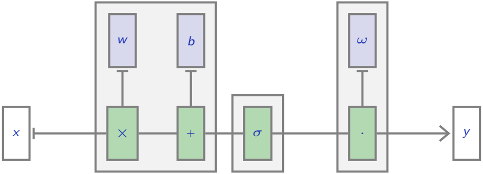
So we can approximate any continuous function

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

with a mapping of the form

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

where $b \in \mathbb{R}^K$, $w \in \mathbb{R}^{K \times D}$, and $\omega \in \mathbb{R}^K$, *i.e.* with a one hidden layer perceptron.



This is the **universal approximation theorem.**

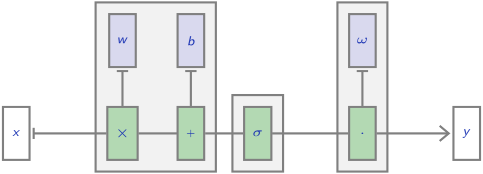
So we can approximate any continuous function

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

with a mapping of the form

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

where $b \in \mathbb{R}^K$, $w \in \mathbb{R}^{K \times D}$, and $\omega \in \mathbb{R}^K$, *i.e.* with a one hidden layer perceptron.



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. We will come back to that later.

The end