Deep learning

3.4. Multi-Layer Perceptrons

François Fleuret

https://fleuret.org/dlc/

A linear classifier of the form

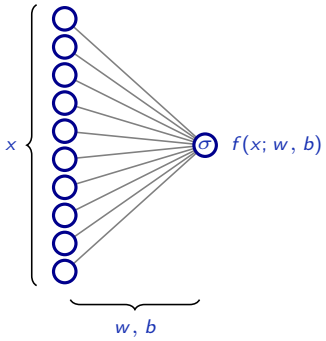
$$\begin{aligned} \mathbb{R}^D &\to \mathbb{R} \\ x &\mapsto \sigma(w \cdot x + b), \end{aligned}$$

with $w \in \mathbb{R}^D$, $b \in \mathbb{R}$, and $\sigma : \mathbb{R} \to \mathbb{R}$, can naturally be extended to a multi-dimension output by applying a similar transformation to every output

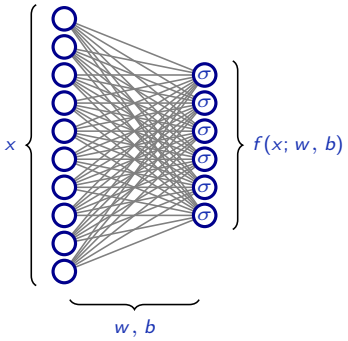
$$\begin{aligned} \mathbb{R}^D &\to \mathbb{R}^C \\ x &\mapsto \sigma(wx + b), \end{aligned}$$

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

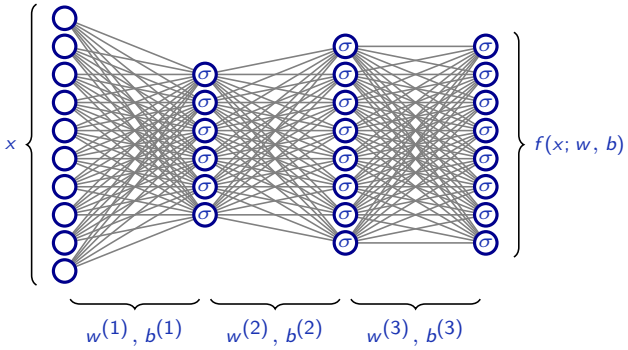
Even though it has no practical value implementation-wise, we can represent such a model as a combination of units. More importantly, we can extend it.



Single unit



One layer of units

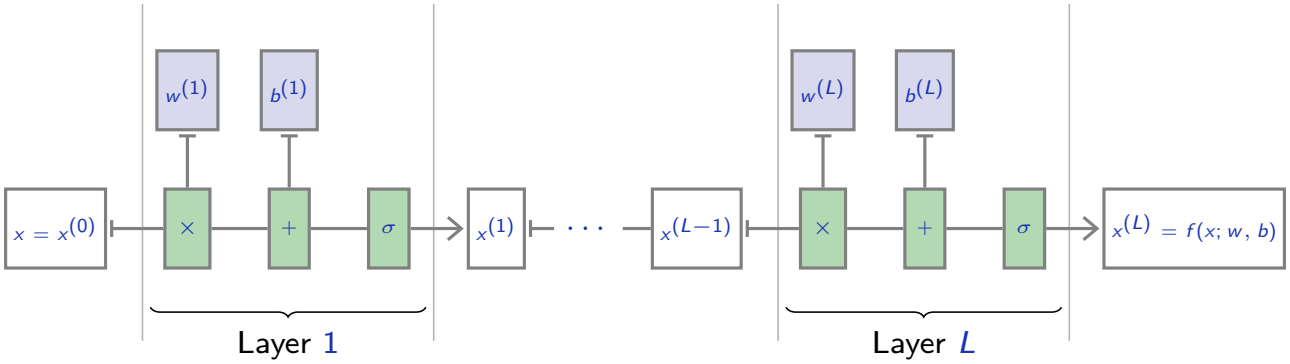


Multiple layers of units

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

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

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



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

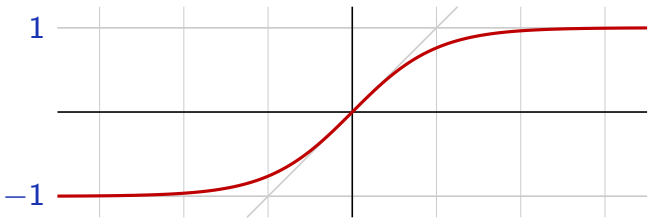
Note that if σ is an affine transformation, the full MLP is a composition of affine mappings, and itself an affine mapping.

Consequently:

The activation function σ should not be affine. Otherwise the resulting MLP would be an affine mapping with a peculiar parametrization.

The two classical activation functions are the hyperbolic tangent

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



and the rectified linear unit (ReLU, Glorot et al., 2011)

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

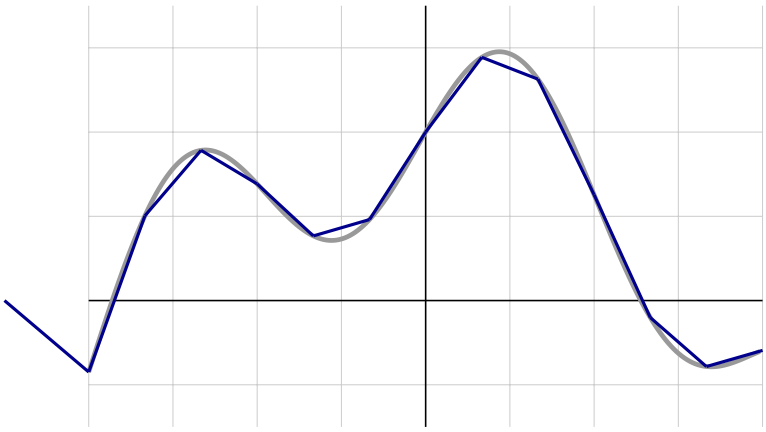
Notes

The hyperbolic tangent was very popular in the nineties. ReLU got popular in the early 2010s, and was one of the reason why deep networks are easier to train.

Universal approximation

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

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



This is true for other activation functions under mild assumptions.

Notes

The universal approximation theorem states that one can approximate a [reasonably regular] function at any precision with a single hidden layer given enough hidden units.
The graph illustrates this for a $\mathbb{R} \to \mathbb{R}$ continuous function using ReLU in the MLP. The blue curve is a sum of several ReLU functions which have been translated and scaled so that their sum is equal to a piece-wise approximation of the target function ϕ shown in gray.
One assumption on the activation function to make such an approximation possible is that the activation should saturate on one side, either $+\infty$ or $-\infty$.

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

We can approximate the sin function with the previous scheme, and use the density of Fourier series to get the final result:

$$\forall \epsilon > 0, \exists K, w \in \mathbb{R}^{K \times D}, b \in \mathbb{R}^K, \omega \in \mathbb{R}^K, \text{ s.t.}$$
$$\max_{x \in [0,1]^D} |\psi(x) - \omega \cdot \sigma(w\, x + b)| \leq \epsilon.$$

Notes

We can use the previous result for the sin function:

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

And the density of Fourier series provides

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

Combining these two approximations provides the result: $\forall \xi > 0$, 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, $\forall x \in [0,1]^D$,

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

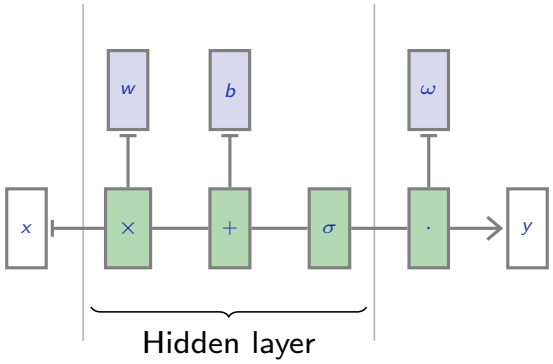
So we can approximate any continuous function

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

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 $\omega \in \mathbb{R}^K$.



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.

So this results states that we can make the **training error** as low as we want by using a larger hidden layer. It states nothing about the **test error**.

Deploying MLP in practice is often a balancing act between under-fitting and over-fitting.

References

X. Glorot, A. Bordes, and Y. Bengio. **Deep sparse rectifier neural networks**. In International Conference on Artificial Intelligence and Statistics (AISTATS), 2011.