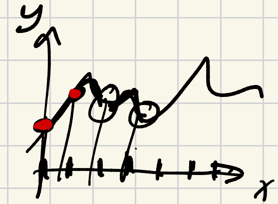
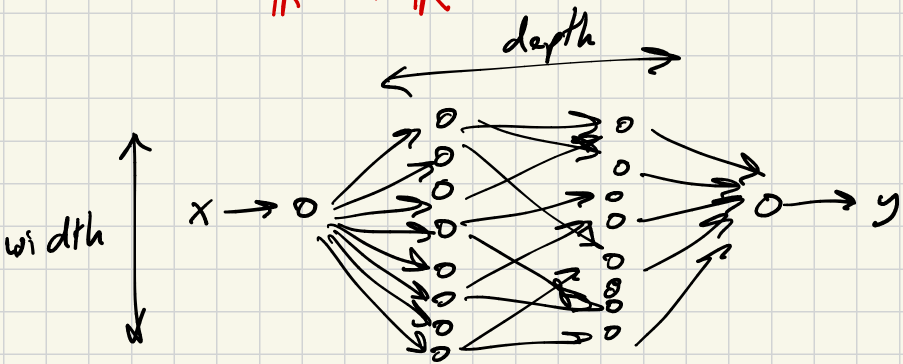
Function approximation with neural networks

$f : \mathbb{R} \to \mathbb{R}$

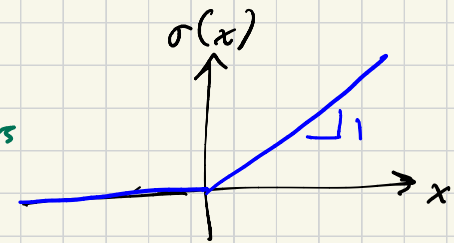
$\mathbb{R}^d \to \mathbb{R}$



#grid points = 10^d



$\sigma : \mathbb{R} \to \mathbb{R}$



$\mathbb{R} \to \mathbb{R}^8 \to \mathbb{R}^8 \to \mathbb{R}$

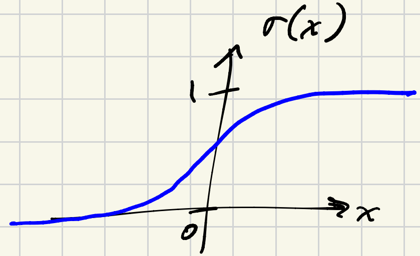
parameters

$x = x_1 \qquad x_2 \qquad x_3 \qquad x_4 = y$

$x_2 = \sigma(A_{12} x_1 + b_1)$	A_{12}	8×1	b_1	8
$x_3 = \sigma(A_{23} x_2 + b_2)$	A_{23}	8×8	b_2	8
$y = x_4 = \sigma(A_{43} x_3 + b_3)$	A_{34}	1×8	b_3	1

MLP

perceptron



Universal approximation theorems

https://en.wikipedia.org/wiki/Universal_approximation_theorem

Single hidden layer, arbitrary width (1990s)

Universal approximation theorem — Let $C(X, \mathbb{R}^m)$ denote the set of continuous functions from a subset X of a Euclidean $\mathbb{R}^n$ space to a Euclidean space $\mathbb{R}^m$. Let $\sigma \in C(\mathbb{R}, \mathbb{R})$. Note that $(\sigma \circ x)_i = \sigma(x_i)$, so $\sigma \circ x$ denotes σ applied to each component of x.

Then σ is not polynomial if and only if for every $n \in \mathbb{N}$, $m \in \mathbb{N}$, compact $K \subseteq \mathbb{R}^n$, $f \in C(K, \mathbb{R}^m), \varepsilon > 0$ there exist $k \in \mathbb{N}$, $A \in \mathbb{R}^{k \times n}$, $b \in \mathbb{R}^k$, $C \in \mathbb{R}^{m \times k}$ such that

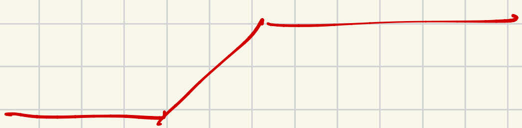
$$\sup_{x \in K} \| f(x) - g(x) \| < \varepsilon$$

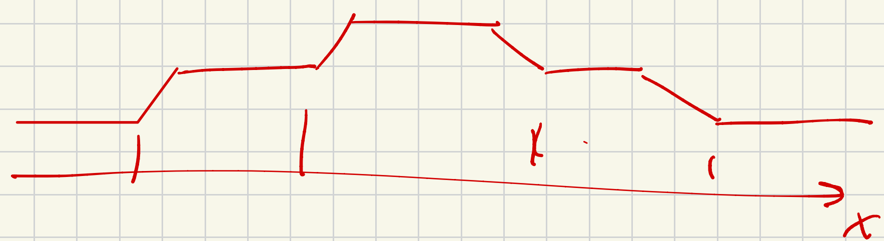
where $g(x) = C \cdot (\sigma \circ (A \cdot x + b))$

Cybenko, G. (1989). "Approximation by superpositions of a sigmoidal function". Mathematics of Control, Signals, and Systems. 2 (4): 303–314. doi:10.1007/BF02551274

Hornik, Kurt (1991). "Approximation capabilities of multilayer feedforward networks". Neural Networks. 4 (2): 251–257. doi:10.1016/0893-6080(91)90009-T

Pinkus, Allan (1999). "Approximation theory of the MLP model in neural networks". Acta Numerica. 8: 143–195. doi:10.1017/S0962492900002919





Fixed width, arbitrary depth (2010s)

Universal approximation theorem *(L1 distance, ReLU activation, arbitrary depth, minimal width).* For any Bochner–Lebesgue p-integrable function $f : \mathbb{R}^n \to \mathbb{R}^m$ and any $\epsilon > 0$, there exists a fully-connected ReLU network F of width exactly $d_m = \max\{n+1, m\}$, satisfying

$$\int_{\mathbb{R}^n} \|f(x) - F(x)\|^p \mathrm{d}x < \epsilon.$$

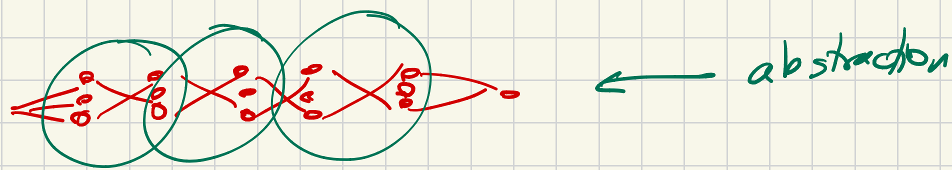
Moreover, there exists a function $f \in L^p(\mathbb{R}^n, \mathbb{R}^m)$ and some $\epsilon > 0$, for which there is no fully-connected ReLU network of width less than $d_m = \max\{n+1, m\}$ satisfying the above approximation bound.

Lu, Zhou; Pu, Hongming; Wang, Feicheng; Hu, Zhiqiang; Wang, Liwei. "The Expressive Power of Neural Networks: A View from the Width". NeurIPS 2017. arXiv:1709.02540

Park, Yun, Lee, Shin, Sejun, Chulhee, Jaeho, Jinwoo. "Minimum Width for Universal Approximation". ICLR 2021. arXiv:2006.08859

R. DeVore, B. Hanin, and G. Petrova, "Neural network approximation", Acta Numerica 30, 327-444, 2021. doi:10.1017/S0962492921000052

deep :



shallow, wide :